Properties and Advances of Probabilistic and Statistical Algorithms with Applications in Finance

Greig Smith

Doctor of Philosophy
University of Edinburgh
2018

Declaration

I declare that this thesis was composed by myself and that the work contained therein is my own, except where explicitly stated otherwise in the text.

(*Greig Smith*)

To my amazing parents and my wonderful fiancée

Acknowledgements

My first thanks goes to my supervisor Gonçalo dos Reis, for his help and advice throughout my thesis. I have certainly learnt a lot and all parts of thesis have benefited immensely due to his knowledge and encouragement. Alongside my supervisor I have benefited from co-authoring with Francisco Bernal, Marius Pfeuffer, Stefan Engelhardt and Peter Tankov. I would like to thank them all for their work which has helped produce and improve my thesis.

On a more personal note, I firstly want to thank my parents and grandparents who have been on this journey with me. Although most of the time it probably sounded like I was speaking nonsense I am grateful for the support they have given. I have also had the joy of meeting some great friends (trapped in the same boat) and in particular I would like to thank Jamie, Jenovah and Luke for some brilliant (mainly non maths) times. My PhD would not have been anywhere near as fun without them.

Finally, one of my biggest thanks goes to my (very recent) fiancée, Shona. You have supported me throughout my whole time at university, been there during some difficult times and always encouraged me to do well and for that I am eternally grateful.

Funding Acknowledgement

Greig Smith was supported by The Maxwell Institute Graduate School in Analysis and its Applications, a Centre for Doctoral Training funded by the UK Engineering and Physical Sciences Research Council (grant [EP/L016508/01]), the Scottish Funding Council, the University of Edinburgh and Heriot-Watt University.

Abstract

This thesis is concerned with the construction and enhancement of algorithms involving probability and statistics. The main motivation for these are problems that appear in finance and more generally in applied science. We consider three distinct areas, namely, credit risk modelling, numerics for McKean Vlasov stochastic differential equations and stochastic representations of Partial Differential Equations (PDEs), therefore the thesis is split into three parts.

Firstly, we consider the problem of estimating a continuous time Markov chain (CTMC) generator from discrete time observations, which is essentially a missing data problem in statistics. These generators give rise to transition probabilities (in particular probabilities of default) over any time horizon, hence the estimation of such generators is a key problem in the world of banking, where the regulator requires banks to calculate risk over different time horizons. For this particular problem several algorithms have been proposed, however, through a combination of theoretical and numerical results we show the Expectation Maximisation (EM) algorithm to be the superior choice. Furthermore we derive closed form expressions for the associated Wald confidence intervals (error) estimated by the EM algorithm. Previous attempts to calculate such intervals relied on numerical schemes which were slower and less stable. We further provide a closed form expression (via the *Delta method*) to transfer these errors to the level of the transition probabilities, which are more intuitive. Although one can establish more precise mathematical results with the Markov assumption, there is empirical evidence suggesting this assumption is not valid. We finish this part by carrying out empirical research on non-Markov phenomena and propose a model to capture the so-called *rating momentum*. This model has many appealing features and is a natural extension to the Markov set up.

The second part is based on McKean Vlasov Stochastic Differential Equations (MV-SDEs), these Stochastic Differential Equations (SDEs) arise from looking at the limit, as the number of weakly interacting particles (e.g. gas particles) tends to infinity. The resulting SDE has coefficients which can depend on its own law, making them theoretically more involved. Although MV-SDEs arise from statistical physics, there has been an explosion in interest recently to use MV-SDEs in models for economics. We firstly derive an explicit approximation scheme for MV-SDEs with one-sided Lipschitz growth in the drift. Such a condition was observed to be an issue for standard SDEs and required more sophisticated schemes. There are *implicit* and *explicit* schemes one can use and we develop both types in the setting of MV-SDEs. Another main issue for MV-SDEs is, due to the dependency on their own law they are extremely expensive to simulate compared to standard SDEs, hence techniques to improve computational cost are in demand. The final result in this part is to develop an importance sampling algorithm for MV-SDEs, where our measure change is obtained through the theory of large deviation principles. Although importance sampling results for standard SDEs are reasonably well understood, there are several difficulties one must overcome to apply a good importance sampling change of measure in this setting. The importance sampling is used here as a variance reduction technique although our results hint that one may be able to use it to reduce *propagation of chaos* error as well.

Finally we consider stochastic algorithms to solve PDEs. It is known one can achieve numerical advantages by using probabilistic methods to solve PDEs, through the so-called *probabilistic domain decomposition* method. The main result of this part is to present an unbiased stochastic representation for a first order PDE, based on the theory of branching diffusions and regime switching. This is a very interesting result since previously (Itô based) stochastic representations only applied to second order PDEs. There are multiple issues one must overcome in order to obtain an algorithm that is numerically stable and solves such a PDE. We conclude by showing the algorithm's potential on a more general first order PDE.

Lay summary

This thesis studies the areas of probability and statistics, although at first it may seem that these are only useful for gambling, as it turns out they have many more applications. For example, people use techniques from probability to model disease spread and techniques from statistics to model earthquake recurrences. However, one of the main applications of these areas, especially probability, has been finance. Since the late 20th century models and algorithms have been developed to describe stock price changes, calculate the chance of a company defaulting, model risk and performance of a hedge fund's portfolio etc.

Although this sounds great, it is important that the models reflect reality and the algorithms are usable (can be calculated in a reasonable time on a computer). One particular problem we focus on is estimating the chances (probability) of a company or country defaulting, this is typically referred to as the risk of the company (country). There are of course some difficulties in accurately modelling this risk. But even when one has a model, obtaining the probabilities of default is still a challenging task. In this thesis we look to improve both the modelling dynamics and the algorithm to estimate the probabilities. One of the other key problems we look into is the case where one has a model, but understanding the dynamics of the model is difficult. An example could be, we have a model for stock price movement and want to know the probability that the price reduces by 10% over the next year, but we can only estimate that using an algorithm (commonly referred to as a numerical technique in this setting). The improvement we look to make here could be cutting computation time (the time taken to run on a computer). These improvements are of course key when one requires a “fast” answer.

Contents

Part I

Introduction

Chapter 1

Estimating Probabilities in Credit Risk Modelling

This thesis is focused on results of probabilistic and statistical algorithms. Such algorithms appear frequently in many aspects of modern life but are also fundamentally important in finance and science more generally. As the problems considered use results and ideas from various areas of mathematics we split the thesis into three distinct parts. In keeping with this theme we also split the introduction likewise, one can view the introductory material here as that for each chapter. In order to keep the thesis as self-contained as possible we also provide a preliminaries chapter at the start of each part where the background material is covered.

How to read this thesis One can read the thesis in a sequential way, however a reader interested in a particular part need not read the other two parts. Namely, each of the three parts are independent of each other, hence one can read, Chapter 1 with Part II, Chapter 2 with Part III and Chapter 3 with Part IV independently. Moreover, the majority of the preliminary material is for the benefit of a reader unfamiliar with the topic and therefore can be skipped by readers more familiar with the material. Some of the notation we use clashes between parts and hence is introduced in the specific part, however, we do introduce several spaces and probabilistic notation in Chapter 4.

Although this thesis considers different areas of mathematics, there is a central theme running throughout; we endeavour to construct and test algorithms for improved convergence and or efficiency.

1.1 Motivation and Background

Before defining our problem in a precise way, let us describe the main motivation for considering this work; credit risk modelling and the 2008 financial crash. This will guide us in framing the problem mathematically.

Credit risk modelling is a key component of any institution which deals with large sums of money. For banks the organisation which sets the regulation for regulatory capital requirement is known as Basel committee on banking supervision (although it is common to refer to this as just the Basel committee and the regulation as the Basel accord). For insurance firms, there is a different regulation, the current is known as Solvency II. Many of the ideas and goals between the two overlap but we will focus on the banking sector. The main aim of the regulator is to develop policies to ensure banks do not take too many risky investments, see [Lüt08, Chapter 2] from some of the history of the Basel accord. Although this was set up in 1988, the financial crisis in 2008 showed up multiple flaws in the safeguards. Many of the assumptions used to model the “riskiness” of an investment simply did not reflect reality. In recent years this has lead to drastic overhauls in the regulation and consequently with some of the simpler models being replaced by more complex ones. Moreover, concepts such as liquidity (ease with which an asset can be sold) that were not truly considered before now feature at the forefront of the regulation[*].

In general, these changes have made the problems mathematically more involved and in some cases highly non trivial. Although the Basel accord governs a large proportion of banking activity, we shall focus on one part, the so-called *risk charge* that a bank must pay, and more specifically estimating the transition probability matrix (TPM) for a company's credit rating. Due to the large sums of money involved, accurate estimation of these probabilities is imperative.

Credit Risk Basics

Let us start by describing key concepts, one can consult texts such as [Lüt08] or [MFE15] for a full discussion on credit risk. Every investment intrinsically carries some form of risk, which brings us to our first definition. *Credit Risk* (or *Credit Worthiness*) is the risk of loss in investment due to some event, generally referred to as a "credit event". That is, the risk that any investment will not be paid in full by the obligor. Traditionally, investments were bonds and the debt holders (investors) that purchased the bonds were concerned that the counterparty (obligor) would default on the bond payment. For reason, credit risk may also be referred to as *default risk*.

Several difficulties can arise when modelling credit risk, the most fundamental of these is lack of data. Default events are relatively rare (especially in a healthy economy) and defaults can be somewhat unexpected. In the event of a default the loss from the portfolio can be large and moreover the size of loss is unknown by the investor before the default occurs. Although default events are rare they are present in every contract that requires a payment obligation. This comes from the fact no future payment is guaranteed, mathematically we can think of this as the probability of not receiving the payment is greater than zero, the so-called, *probability of default* (PD). Clearly, the PD is not constant in time, the longer the time horizon the greater the PD. Typically PDs are quoted over the one year time horizon, that is the probability that the obligor defaults within one year.

Even at this point we have a difficulty to overcome, how does one assign a default probability to every borrower in a bank's credit portfolio? One approach is to calibrate default probabilities from market data, this is the concept used by KMV where they calculate the EDF (Expected Default Frequencies), see [MFE15, Chapter 8] for further details. Another method to compute default probabilities is through the credit risk inherent in credit spreads of certain traded products, commonly *credit default swaps* (CDS), see [MFE15, Chapter 9] . Finally, and also most common is using credit ratings either assigned through an external rating agency such as *Moody's investor service* or *Standard & Poor*, or indeed by the bank's on internal rating system. Closely related to default probabilities is that of *credit migration* and *migration risk*, which is the loss due to an asset changing its PD, typically through a change in rating. We shall focus heavily on credit ratings.

Returning to an earlier point, an obligor defaulting only implies they cannot match the full demands of their debts. Hence, an obligor defaulting does not imply that all the money from the investment is lost, only a proportion. We define the *exposure at default* EAD of an obligor as the portion of the exposure (investment) which is lost in the event of default. We further define the *recovery risk* as the uncertainty about the proportion of loss in the event of default. In the event of recovery (some amount being paid back) the uncertainty is in the amount of money received after default, see [Tas04] or [MFE15, Chapter 8] for example.

Credit Ratings

Credit ratings play a key role not just in the calculation of a bank's capital charge (amount of capital a bank must hold) but also are typically a requirement for corporations wishing to issue bonds. There are different agencies which provide firms with a rating and the credit rating the agency gives a company determines in some respect the financial health of the company (or country). Ratings are typically given in terms of letters, for example AAA, AA, A, BBB, BB, B, C, D (although it varies between agencies) with, AAA the best (safest), C the worst (riskiest) and D to imply the firm has defaulted. It is also standard for banks to use their own internal

[*] The current regulation can been found on the Basel committee website *https://www.bis.org/bcbs/*.

ratings system (see [YWZC14]). Rating grades are typically put into two categories, "investment grade", which for example is BBB and better and "speculative grade" which are the lower rated firms. For an overview of the 'science' involved in the rating procedure see [Can04].

The main object of interest in this work is the so-called annual *Transition Probability Matrix* (TPM), it is a stochastic matrix which shows the migration probabilities of different rated companies within a year. Rating agencies produce these annually. It is possible that such matrices are not initially stochastic due to company mergers or rounding for example. However, they can be renormalised by methods as described in [KS01] and, as argued in [BDK$^+$02], renormalising non rated companies across all ratings is indeed the industry standard.

One problem considered here is that a TPM encases transition probabilities over a 1-year time frame and often in practice one needs a 3 month or 10 day transition matrix for which probabilities of default are lower than those in the TPM. Therefore one wants to accurately estimate the sub-annual matrix given the annual matrix. In the Basel proposals, Basel 3 [Sup13, p.3] a large part of the risk charge will be measured using ES (expected shortfall), which (as shown in [CDS10]) is extremely sensitive to a small shifts in probabilities. Therefore, accurate and consistent estimation is essential in the calculation. Moreover, with the perspective of the *IFTR 9* regulation (most recent regulation), one needs to better estimate the related probabilities of default (PD) since for a company whose risk profile changes significantly, one needs to assess its risk throughout the bond's lifetime.

Credit ratings are of course not without their own issues, in theory the rating given to a company should reflect the risk of that company. However, due to trends such as economic fluctuations there are reasons one would want to know the risk over a longer period of time, which is also viewed as the more prudent view point. This longer time view is known as *through the cycle* (TTC), while the short term risk is known as *point in time* (PIT). There are different arguments for which is better, typically rating agencies use TTC because they provide more stable ratings, although are clearly less useful for short term investments. We shall not go into details here, however, one can consult [IJJK14] or [Cou08] for further discussion on this topic.

The final point to make on credit ratings is there are essentially two levels of information. As mentioned previously it is common for rating agencies to produce annual (empirical) transition matrices. However, this data is anonymous, namely, one cannot track the movement of an individual company. Consequently, all companies in the same rating must be treated equally which corresponds to the so-called *Markov assumption*. This implies that the best we can do is model these transition probabilities in a Markovian way, we will come back to this in detail in Chapter 6. It is possible (at a cost) to buy the full ratings data set which allows one to track the movements of each company. Provided one has access to this data set it allows use to consider more general (non-Markov) models, again we shall return to this in Chapter 7.

Risk Charge and Regulation

The final part of this discussion involves regulation and associated with that is the risk charge. For the benefit of a reader unfamiliar with the area we will try to avoid banking terminology, however, an interested reader can consult texts such as [CCP12] for precise details. The main motivation for estimating the probabilities comes from the calculation of *risk charges*. A risk charge is an amount of capital a bank must hold back (not invest) to cover against large losses. As mentioned earlier, new regulation was introduced post 2008 to improve the previous regulatory issues surrounding the amount of capital a bank is required to hold. It should be mentioned here that regulation can be updated frequently, the regulation that came in post 2008 crash, Basel 2.5 has been replaced recently by Basel 3. Because from our perspective Basel 2.5 is simpler to explain and in either framework, probabilities (mainly of default) still need to be estimated. We stick to describing some of the changes between Basel 2 and Basel 2.5 and mainly use this framework when testing models.

One of the key changes from Basel 2 to Basel 2.5 is the charge known as the *IRC* (Incremental Risk Charge) component, but in order to understand why IRC came into existence we must first look at the 2008 financial crisis. Although banks had a risk charge at the time of the crash it was easily overwhelmed and this comes down to serious underestimation of risk at both the level of the banks and the Basel committee. Mortgage backed securities being a classic example of where risk was severely underestimated.

The committee deals with various aspects of banking, however we will be focusing on the non-securitised (the investment isn't secured against an asset, for example a mortgage is secured against a property) capital requirements associated to pillar 1 of the regulation. A capital requirement (or risk charge) is money which a bank is forbidden to invest, that is, it must hold money at a central bank. The level of the capital held is calculated based on the risk of a bank's investment, such that it doesn't go bankrupt under a stress scenario. As mentioned above, unfortunately the regulation during the crisis was based on assumptions that were not true of a stressed market. One assumption in particular was that a bank could sell certain positions (investments) within 10 days, these are referred to as liquid, see [Sup13, pg.4]. As it turns out, many of these positions that were thought to be liquid became illiquid in a stressed market. Furthermore, the measure of loss was purely concerned with the probability of default, and is not interested in so-called credit migration, namely, when a position changes its rating (quality) [Sup09, pg.1]. Since higher quality positions are more expensive (lower return), a rating migration will change its value and if the bank cannot sell the position there is more scope for it to further decrease in quality. In fact due to these flaws, the majority of money lost from banking organisations came from losses of this form [Sup09, pg.1]. In some cases this was enough to bankrupt the bank.

With the above in mind, one of the main goals of Basel 2.5 was to make more robust estimates on a bank's ability to trade in a stressed market, see [Sup09] for a full account. In essence this meant exchanging the 10 day selling period to a minimum three month selling period, a so-called *liquidity horizon*. Also, to account for the potential loss in value due to a rating downgrade, the credit migration risk. Therefore under this changed regulation a bank must calculate the risk to any of its investment over a three month period, where risk incorporates default and migration risk.

The latest regulation (Basel 3) concentrates even more on this idea of liquidity and appropriate *stress-testing* of banks, namely, how well would a bank perform under a theoretical stress scenario. As mentioned previously the precise requirements from different regulations can change but at the core one must always model and hence estimate quantities such as probability of default.

Key Questions.

Now that we have detailed the background on the problem, let us define our key questions.

- How do we convert this problem into a framework that allows us to use the tools from mathematics? How does this vary under the two levels of information mentioned above, namely, annual and individual transition results?
- In the case of less information, what methods can be used and is there a "best" algorithm to use? Moreover can we obtain error estimates from the algorithm using the Fisher information matrix?
- In the case of obtaining more information, we can use more general, non-Markov models. How does this new model effect the rating transitions and in particular probabilities of default?

We shall answer these questions in Part II, however, let us discuss here how the mathematical framework for credit risk can be introduced.

1.2 Mathematical Framework

Now that one has some understanding on the problem and how it is used in a wider context let us express the problem mathematically and develop a strategy to tackle it.

1.2.1 Estimating a Markov Chain with missing data

As previously mentioned, a bank can have access to the overall transitions of all companies over the year (the TPM) and not to the set individual company movements (which are more expensive). As we also discussed, companies can leave the system and this would make the TPM not a stochastic matrix, however, the matrix one obtains from the ratings agency is reweighted in such a way that it is. We do not go into how valid a method is, the reader can consult [BDK$^+$02] for such discussion.

There are three points to make here, firstly the rating assumption automatically yields a Markov structure (see Section 5.1 for details on Markov chains), that is, we take every company in every rating to have the same risk. Hence the company's default probability is completely determined by its current rating so previous ratings do not change this. Secondly, we only receive annual data, this implies we have a missing data problem e.g. a company can jump from rating $A \to B \to A$ in one year and we would observe no movement. Finally, our TPM is empirical, therefore there are possible transitions that we do not observe, but we still want to estimate a probability of such an event.

Hence by taking a company's rating as X, the goal is to estimate the Markov chain (stochastic matrix) governing X which implies its probability of rating transitions and in particular the probability of default. There are two types of Markov chain we may use, we can either view this as a discrete Markov process or a continuous one, a so-called continuous time Markov chain (CTMC). We focus on the CTMC approach (we give some justification for this in Remark 1.2.2), such models are determined by a so-called generator matrix Q. When one has a generator matrix it is simple to obtain the stochastic matrix for X over any time interval t by taking[†] e^{Qt}. Hence, we wish to estimate Q to understand the risk of each rating.

The agent only has the annual empirical TPM, say P, and using this data wishes to fit a CTMC. That is we want to find a generator matrix Q, such that e^Q is "close" to P. Although one can often find a matrix Q such that $e^Q = P$, typically this is not a generator, hence e^{Qt} will not be a stochastic matrix for all t. This makes the estimation non-trivial and is referred to as the *embeddability* problem. It is discussed in great detail by [IRW01] and, for more of the mathematics and many of the existing results on the embeddability problem, we point the reader to [Lin11].

Theorem 1.2.1. *Let P be a transition matrix with entries $(P_{ij})_{i,j=1,\dots,h}$, and suppose that either, (a) $\det(P) \leq 0$, or (b) $\det(P) > \prod_i p_{ii}$, or (c) there are states i and j such that j is accessible from i, but $p_{ij} = 0$.*

Then, there does not exist an exact generator for P.

Since every empirically observed TPM will satisfy one of the conditions Theorem 1.2.1, typically condition[‡] c). It implies one cannot simply take $\log(P(1))$ (where $\log$ is the matrix logarithm) to obtain Q.

Several approaches exist to tackle this estimation problem [KS01], [IRW01], [TÖ04], [BS05], [Ina06], [BS09], either using deterministic algorithms (e.g. diagonal or weighted adjustment, Quasi-optimization of the generator) or statistical ones (Expectation-Maximisation (EM), Markov chain Monte-Carlo (MCMC) ones), see Section 6.2. We focus on the Expectation-Maximisation algorithm of [BS05] for CTMCs and allow for an absorbing state. We will give full details on these methods in Chapter 6.

Remark 1.2.2 (Continuous over Discrete). *We focus purely on continuous over discrete time models. Continuous time algorithms yield robust estimators while the discrete ones do not, with robustness understood in the following sense: from $P(1)$ estimate $P(0.5)$ and $P(0.25)$. From $P(0.5)$ estimate $P(0.25)$ again. Continuous algorithms yield the same $P(0.25)$, discrete algorithms (in general) will not.*

In theory, a discrete Markov process is more accurate i.e. a company would not get rerated more than once in a single day. However this would lead to an extremely computationally heavy model since all possible combinations of paths would need to be considered. Therefore the expressions on can derive from continuous models are easier to work with.

[†] Since Q is a matrix, e is the matrix exponential.

[‡] There has never been an observation of a AAA company defaulting over a one year horizon, though in theory it could happen.

1.2.2 Removing Markovian Assumption

Credit rating models within the Markovian framework are handy both from a theoretical and numerical perspective. As shown in [LS02] (with access to the “full” dataset), rating transitions exhibit non-Markov features. Namely, an obligor that has recently downgraded into a certain rating is more likely to downgrade further than other obligors currently in that rating. Such an effect is referred to as *rating momentum*; momentum may also appear in upgrades, however, it is not as apparent. Main (non-Markovian) effects include *rating drift* (or *momentum*) [AK92] and [LS02], *rating stickiness* [MFE15] and specific rating agencies’ policies (see [CH01] and [Løf05]). [NPV00] highlight non-Markovian patterns in transition probabilities for ratings and discuss their dependence with respect to the parameters like industry, domicile and business cycle. However, of these effects rating momentum is the most important to capture and what we look to model here.

The rating momentum effect has a non-negligible bearing on the risk attributed to a portfolio as it makes defaults of investment grade bonds likelier than in the standard Markov set up. [Cou08, p.8] report on the temporal span of the rating drift (for a certain Standard & Poor’s database) and its mean reversion. Over longer horizons the non-Markov effects such as momentum become more pronounced, i.e. have a larger impact on transition probabilities. At a practical level the IFRS9 regulation requires knowledge of risks on rating migrations over longer horizons, hence these effects can significantly change the results. When one has access to the full data set it is possible to create models that capture non-Markov effects and this is one of our contributions. The proposed model captures the momentum effect which leads to an interesting result, whereby we find the purely Markov model underestimates default risk in investment grades, but overestimates the risk in some speculative grades. We discuss this further in Section 7.2.

Potential Non-Markov Models

Different models have been introduced in the past to incorporate non-Markov phenomena. We briefly overview some of these works here.

There are several works such as [MW07], [KP07] and [FBSN16], which look at using the observed transition matrix to model the movement of companies. However, these works are mainly concerned with portfolio and industry wide modelling to capture systematic effects such as contagion. This is different to the obligor’s dependence on its own history, which we wish to model here.

Extended State Space and Mixture Models. [CHL04], attempt to take non-Markovian effects into account while keeping some Markovian structure. The idea is to extend the state space to include $+$ and $-$ states, for example when a company downgrades from A to B, it goes into state B^{-}, which has higher probabilities of downgrades than B. Similarly if the company moves from B to A it goes into A^{+} which has smaller probabilities of downgrades than A. This allows us to keep the Markov assumption, however, we must calibrate many more parameters and we do not observe a company belonging to the excited or non-excited state. Therefore when successive transitions occur, it is unknown whether the company was in the excited or non-excited state. Hence calibrating an intensity between excited and non-excited states seems impossible. One could navigate around this by assuming excited states do not jump to non excited states, but this is against empirical evidence of momentum reducing over time, see [Cou08] for example.

[DJM16] apply a semi-Markov model to capture the observed effect that companies move from states not following an exponential distribution. However, they still rely on the Markov transition structure. Hence they need to expand the state space in order to include momentum. Related to this approach is [FS08], where the authors use two different time homogeneous CTMC generator matrices (thus making transitions non-Markov), however, it does not capture momentum since the jump itself is Markov.

Hidden Markov Model (HMM). A different idea is to use a hidden Markov model (HMM) (see [CMR05] for a complete account). The HMM approach to credit risk can be traced back to the work of R. Elliot and is described in [Kor12]. In rough, the approach considers two processes (X, Y), the observed (published) credit rating Y and the “true" credit rating X which is unobserved, i.e. hidden. The paradigm is, credit ratings are assumed to be “noisy” observations and not the true representation of the credit risk. The goal is then to use Y to make inference on X. In such a setup if one considers the noisy observation and the true rating as correlated, then rating momentum can be added into the model. Although this approach has some benefits, the work appears to be constrained to the discrete time case.

Hazard Rates, Point Processes and self-exciting Marked Point Processes. Let us start by discussing Hazard rates, the main work in this area for credit ratings is given in [KLM08]. An extensive work bringing hazard rate methodologies to the estimation of probabilities of default can be found in [Cou08] (and references therein). The paradigm is that each company has a corresponding hazard rate (a parameter), in this hazard rate one can encode various factors such as momentum for example. The issue with [KLM08]'s methodology is that they must calibrate parameters for each of the various transitions with the extra variables to obtain the probabilities of these transitions. This however, increases the model's complexity greatly. Our goal is to present a model as parsimonious as possible that captures rating momentum.

Our approach relies on *point processes* that are dependent on their own history, so called *self-exciting* processes (see [DVJ03], [DVJ07]). Point processes are generalisations of Markov processes and hence a natural choice for our model. One of the most satisfying results of using point processes though is that one can capture rating momentum by adding only a small number of parameters. The most common example of a self-exciting process is a Hawkes process. These processes appear in other areas of mathematical finance, such as limit book orders and high frequency trading [BMM15], however, they have not been fully utilized in credit transitions. A Hawkes process can be thought of as a counting process (similar to a Poisson process) which in one dimension has an intensity λ_t of the form (see [DZ13]),

$$\lambda_t = \mu + \int_0^t \phi(t-s)\mathrm{d}N_s\,, \tag{1.2.1}$$

where N is a counting measure and denotes that an event has occurred (this will be a rating change in our case), and ϕ is the impact on the intensity and allows the intensity to depend on previous events. By setting $\phi = 0$ the Hawkes process reduces to a Poisson process. A common choice for ϕ is the so-called exponential decay, namely $\phi(t-s) = \alpha\beta\exp(-\beta(t-s))$ with $\alpha, \beta > 0$. Functions of this form are useful since the event's influence on the intensity weakens as time progresses, hence we can account for momentum reducing over time (agreeing with the findings of [Cou08]).

Hawkes processes in this form are not fit for our purposes since we require different changes to intensity dependent on whether it is an upgrade or a downgrade. Further we require the baseline intensity μ must depend on the current state. Such processes are referred to as *marked point processes*, since to each event observed one assigns a *mark* to indicate the type of event, see [DVJ03, Chapter 6.4]. We discuss this further in Section 5.4 and 7.2.

1.2.3 Contributions

Part II contains many results in the field of credit risk.

Firstly in Chapter 6 we conduct a series of tests on the various algorithms in the literature to estimate Q. We find that the EM algorithm is the best choice for this problem in terms of accuracy and computational efficiency. Moreover, under an extremely mild assumption for credit risk we strengthen the convergence result of the EM algorithm, this ensures the EM converges in parameter space, hence the values it returns are stable from one iteration to the other, see Theorem 6.1.8. This is of course key since it implies a robustness about the algorithm that was not previously known. Moreover in Theorem 6.1.12 we derive a closed form expression for the error in the estimated Q via the Fisher information matrix. This is a substantial improvement on the previous numerical approaches which are slow and due to the often small parameter values unstable. Finally our closed form expression for the error in Q allows us to further obtain a closed form expression for the error in the transition probabilities, through the so-called delta method, see Theorem 6.1.14. This has never been fully considered before, however, it is highly useful for financial institutions since the errors are easily interpretable.

In Chapter 7, we look to remove the Markov assumption, that is we do not view every company in the same rating to have the same risk. This effect was observed in [LS02] when the authors provide evidence showing that a company which has been recently downgraded is more likely to be further downgraded, so-called *momentum*. However, there has not been a simple model produced that captures this effect. We present a *marked self-exciting point process* model which one can view as a generalisation of a Markov model which allows for rating momentum. In order to calibrate this model (and any non-Markov model for that matter) we require access to the "full" data set and this will be discussed further in Chapter 7. Our model has many appealing features such as only requiring four extra parameters (compared to the ≈ 50 already in the Markov model). Moreover, as shown in Figure 7.1 we observe an interesting consequence of using a non-Markov model, although default risk for investment grade companies are more likely with the momentum effect, speculative grade companies are viewed to be less risky. This indeed has implications for banks and regulators when calculating risk charges. To substantiate our claim we test the Markov and non-Markov model against empirical observations from the Moody's proprietary corporate credit ratings data set and found that the Markov model does appear to overestimate risk on some speculative grades and underestimate on investment grades.

Chapter 2

McKean-Vlasov SDEs

2.1 Motivation and Background

The algorithms mentioned in Chapter 1 are statistical algorithms, that is, one has data and a model, but is interested in estimating (fitting) parameters to the model. Of course there are other problems which require algorithms, such as estimating the solution to an equation or approximation of an integral etc. However, in this setting they are typically referred to as numerical techniques and we shall use the terms interchangeably throughout. The judge of quality of the algorithm (numerical technique) is its accuracy in approximation and its computational complexity.

This chapter focuses on simulation and estimation of McKean Vlasov Stochastic Differential Equations (MV-SDEs). MV-SDE are stochastic differential equations where the coefficients of the SDE depend on the law of the solution, typically written in the following form:

$$\mathrm{d}X_t = b(t, X_t, \mu_t)\mathrm{d}t + \sigma(t, X_t, \mu_t)\mathrm{d}W_t, \quad X_0 = x, \tag{2.1.1}$$

where μ_t denotes the law of the process X at time t i.e. the pushforward measure $\mu_t = \mathbb{P} \circ X_t^{-1}$, and W is a standard Brownian motion (under $\mathbb{P}$). The motivation to study these equations originated from statistical physics. That is, one can consider a large number, N of weakly interacting particles whose dynamics are given by, for $i \in \{1, \dots, N\}$, with $X_0^{i,N} = x_0$

$$\mathrm{d}X_t^{i,N} = b\Big(t, X_t^{i,N}, \mu_t^{X,N}\Big)\mathrm{d}t + \sigma\Big(t, X_t^{i,N}, \mu_t^{X,N}\Big)\mathrm{d}W_t^i, \quad \mu_t^{X,N}(\mathrm{d}x) := \frac{1}{N}\sum_{j=1}^N \delta_{X_t^{j,N}}(\mathrm{d}x) \tag{2.1.2}$$

where $\delta_{X_t^{j,N}}$ is the Dirac measure at point $X_t^{j,N}$, and the Brownian motions (BM) $W^i, i = 1, \dots, N$ are independent. It is straightforward to see that motion of particle i depends on all other particles j, however, the dependence on any individual particle becomes weaker as N becomes larger. For large N, (2.1.2) is a difficult system to study, however, as shown in [Szn91], [Mél96] and [Car16] under appropriate Lipschitz style assumptions one can obtain the following pathwise convergence result,

$$\lim_{N\to\infty} \sup_{1\le i\le N} \mathbb{E}\left[\sup_{0\le t\le T} |X_t^{i,N} - X_t^i|^2\right] = 0$$

where X^i, the solution of (2.1.1) driven by the Brownian motion W^i. Such a result is typically referred to as *propagation of chaos*, it implies that (2.1.1) is the large N (or asymptotic) limit of (2.1.2), therefore one can study (2.1.1) to understand the particle system. As it turns out we will need to generalise this result, hence we will come back to it in more detail later.

Although MV-SDEs have their roots in statistical physics, they have many other applications such as fluid dynamics and weather prediction, see [BBC$^+$10] and modelling neuron activity in the brain, see [DIR$^+$15]. Recently there has been a large increase in interest from economics and finance using MV-SDEs to model systems with a large number of competing agents. Classically so-called stochastic differential games involved a set of N agents following some dynamics and was written as a stochastic control problem. One is then interested in finding the agent's "optimal" action, from a given set of *controls*. The standard example is when all agents are treated equally and finding the optimal amounts to determining a Nash equilibrium. There is a vast array of work on stochastic games and stochastic control and one can consult texts such as [Car08], [Pha09], [BLR17], [Car16] for various examples. One should note however, these models rely heavily on so-called BSDEs (Backward Stochastic Differential Equations) and FBSDEs (Forward BSDEs), this is a fascinating but complex subject and we do not discuss it here, although we will give further details in Part IV.

Similar to above, one may be interested to know or it may be of benefit to instead of considering an extremely large (but finite) number of agents (which all act in a statistically similar way), approximate the dynamics by taking the number of agents to infinity. By making the approximation of an infinite number of agents, this reduces the complexity of the model and allows for more tractable results. Again the problem is to solve for a Nash equilibrium. There are two ways in which one can do this, firstly optimise with a fixed measure, then take the limit $N \to \infty$ and solve for the law (find the fixed point where the measure coincides with the law), this is referred to as a *mean-field game*, as developed by Lasry and Lions in 2006, see [GLL11] and [CD17a, Chapter 3] for further details. The second way is the reverse order, i.e. take the limit $N \to \infty$ first with a fixed control, which then yields a so-called optimisation problem over control dynamics of McKean-Vlasov type, then solve the optimisation problem, one can consult [CD$^+$15] and [CD17a, Chapter 6] for results in this area. As it turn out the two approaches lead to different results as discussed in [CDL13]. Such approaches can be used to model portfolio strategies of jealous investors or runs on a bank, said differently, these are useful tools when one looks to take behaviour into account (herd behaviour for example) which is of clear importance in finance and economics. Outside of finance mean field games have also recently been used to model optimal energy consumption and storage, see for example [MAT18]. For a full account on both methods and their applications the reader can consult [CD17a] and [CD17b].

Although mean-field game type problems constitute the main application of MV-SDEs in finance, there is other work which relates MV-SDEs and their connect with SPDEs to model mortgage backed securities (MBS), see [AHL18]. MBS are where a bank "pool" a large number of its mortgage customers and sell the mortgage repayment stream to investors, essentially creating a bond. During the financial crisis in 2008 MBS are often looked upon as one of the key drivers, mainly due to the size and complexity of the mortgage pools making it difficult to understand the risk, but also the fact that many banks were putting so-called *sub prime* mortgages in their MBS making them riskier investments than initially thought. What makes these sorts of asset ideal for modelling using McKean-Vlasov type equations is, it is very easy to capture effects such as contagion via the interaction. Another more direct use of MV-SDEs is in stock prices with dividends, see [Bañ18]. The idea here is that dividends are paid based on an expected stock price, rather than just the realised.

Further references and results. Here we shall only consider standard MV-SDEs and numerical techniques for them. Although there has been a great deal of work extending results and ideas from standard SDEs to the McKean-Vlasov setting which we provide some references for here. There has been work on linking such MV-SDEs to solutions of PDEs (Fokker-Plank) and SPDEs, see [DIR$^+$15], [CX10], [BLP$^+$17] and [CM17b] and references therein. For differentiability and Malliavin calculus results on MV-SDEs, the key main difficulty is how to "differentiate" on the measure component. The idea on how one can do this was solved by Lions (see [Car10] for details), and subsequently work has been carried out by [BLP$^+$17], [CCD14], [CM17b] and [Bañ18] and references therein. Finally, [BLP09] and [BDL$^+$09] developed the McKean-Vlasov type extension to BSDEs, a so called mean field BSDEs (or FBSDEs), this work is critical in order to establish the results on mean-field games as discussed above.

Key Questions.

Let us now set out the questions we wish to address.

- How does one simulate a MV-SDE and how does this compare to the simulation of standard

SDEs?

- What are the current limitations surrounding the simulation of MV-SDEs and can one look toward SDEs as inspiration to improve the simulation? That is, can we simulate MV-SDEs with superlinear growth?
- How does the measure dependency affect the computational cost and again can we take inspiration from SDEs to build an improved algorithm variance reduction? In particular can we perform *importance sampling* with MV-SDEs.

We address these questions in Part III, however, let us discuss the first question here.

2.2 Current Results on Simulation

In order to motivate why these questions are important, let us discuss the differences in simulation between SDEs and MV-SDEs. Despite MV-SDEs having recently had an increase in their applications and popularity, the simulation of these equations is still more challenging compared to standard SDEs. This remains one of the main issues regarding the use of MV-SDEs. This stems from the fact that (in general) there is no clear, easy alternative way to approximate the law using stochastic techniques. Therefore and somewhat perversely, although (2.1.1) is typically more beneficial from the theoretical point of view, it is not particularly useful from the numerical standpoint. For simulation purposes, when the law of the SDE is a priori unknown, one typically reverts back to (2.1.2).

Using Lipschitz (or stronger) assumptions, one of the earliest works in the field is a convergence result in [Szn91] showing convergence of the interacting particle system to the MV-SDE. Following that [BT97], [KHO97] and [Bos04] worked on the convergence rate of a Euler type interacting particle scheme. There was not much work done after this until recently where three different approaches to the problem have been proposed in order to make the simulation more efficient [GP18], [CM17a], [HAT18] and [STT17] among others. Firstly, [GP18] navigate around the problem of approximating the law of the MV-SDE via a large particle system by writing an approximation scheme to the law directly using PDEs. Although this requires an additional step we can use the same law for all particles, which removes a large part of the computational cost. Secondly, [STT17] and [HAT18] apply Multilevel Monte Carlo techniques to improve the simulation of MV-SDEs. That is they develop a Multilevel Monte Carlo scheme which has a smaller weaker error. Finally [CM17a] use *cubature* methods as developed in [LV04], such methods have been shown to be effective alternatives to Monte Carlo and again in this setting eliminate the need for a particle system. Although these papers have shown superior convergence to that of standard Monte Carlo with interacting particles, they all rely on sufficiently smooth coefficients and all are required to be Lipschitz. Moreover, they all require scalar or first order interaction with the law. In this chapter we focus on schemes that do not require differentiable and Lipschitz-type coefficients and show the enhancements one is able to make in this setting.

To see why MV-SDEs are much harder to deal with than standard SDEs let us consider the following example. If one is interested in estimating a quantity such as $\mathbb{E}[G(X_T)]$, then the system (2.1.2) is a system of ordinary SDEs. In general one cannot simulate SDEs exactly either and hence requires a numerical scheme, such as the Euler scheme (under sufficiently nice coefficients). We partition the time interval $[0,T]$ into M steps of size $h := T/M$, we then define $t_k := kh$ and recursively define the particle system for $k \in \{0,\dots,M-1\}$ as,

$$\bar{X}^{i,N,M}_{t_{k+1}} = \bar{X}^{i,N,M}_{t_k} + b\Big(t_k, \bar{X}^{i,N,M}_{t_k}, \bar{\mu}^{X,N}_{t_k}\Big)h + \sigma\Big(t_k, \bar{X}^{i,N,M}_{t_k}, \bar{\mu}^{X,N}_{t_k}\Big)\Delta W^i_{t_k},$$

where $\bar{\mu}^{X,N}_{t_k}(\mathrm{d}x) := \frac{1}{N}\sum_{j=1}^N \delta_{\bar{X}^{j,N,M}_{t_k}}(\mathrm{d}x)$, $\Delta W^i_{t_k} := W^i_{t_{k+1}} - W^i_{t_k}$ and $\bar{X}^{i,N,M}_0 := X^i_0$. Let $X^{i,N,M}$ be the i-th component of the solution of (2.1.2), discretised on $[0,T]$ over M steps. The quantity of interest, which, in our case is $\theta = \mathbb{E}[G(X_T)]$, will then be approximated by the Monte Carlo estimator

$$\theta \approx \hat{\theta}^{N,M} = \frac{1}{N}\sum_{i=1}^N G(X^{i,N,M}_T).$$

Unlike standard SDEs which are affected by two sources of error (statistical and discretisation), the MV-SDE approximation is affected by three.

- The statistical error, that is the difference between $\hat{\theta}^{N,M}$ and $\mathbb{E}[G(X^{i,N,M})]$.
- The discretisation error (bias), that is, the difference between $\mathbb{E}[G(X^{i,N,M})]$ and $\mathbb{E}[G(X^{i,N})]$.
- The propagation of chaos error of approximating the MV-SDE with the interacting particle system, that is, the difference between $\mathbb{E}[G(X^{i,N})]$ and $\mathbb{E}[G(X)]$.

The discretisation error of ordinary SDEs has been analyzed by many authors, and it is well known that, e.g., under the Lipschitz assumptions the Euler scheme has weak convergence error of order $\frac{1}{M}$ (see [KP11, Chapter 14]). It is of course well known, the standard deviation of the statistical error is of order of $\frac{1}{\sqrt{N}}$ (see [Gla13, Chapter 1]). There has also been some work detailing the error from the propagation of chaos as a function of N, essentially for G and X nice enough the weak error is also of the order $\frac{1}{\sqrt{N}}$, see for example [KHO97] and [Bos04] for further details.

Although all three errors are of interest, we shall concentrate on the discretisation and statistical error.

2.2.1 Non Lipschitz Simulation

It is well documented that for standard SDEs, the Euler scheme diverges when one leaves the realm of Lipschitz (or linear) type growth and one considers monotone or one-sided Lipschitz growth in drift (see [HJK11]) for example, for all $t \in [0,T]$,

$$\langle x - x', b(t,x) - b(t,x') \rangle \leq L|x - x'| \, ,$$

where $\langle \cdot, \cdot \rangle$ is the standard scalar product and $| \cdot |$ our norm. This is in spite of the fact that the SDE is known to have a unique strong solution, see [Mao08, Theorem 2.3.5]. This is not ideal since many SDEs one wishes to use in practice have coefficients that only satisfy these more general conditions, such as the stochastic Ginzburg Landau equation,

$$\mathrm{d}X_t = \Big(\Big(a + \frac{\sigma^2}{2}\Big)X_t - bX_t^3\Big)\mathrm{d}t + \sigma X_t \mathrm{d}W_t, \quad X_0 = x,$$

where a, b and σ are positive constants, see [Tie13] for more discussion on this model and its applications. There are also many such SDEs that are commonly used in finance which have nonlinearly growing coefficients, see [CJM16] for several such examples. One possible way around this is to modify the drift such that it remains bounded but the scheme still converges to the true, there are two such methods *taming* as developed in [HJK12] (see Section 8.3 for further details) and *truncation*, see [CJM16].

It has also been shown that MV-SDEs with non Lipschitz growth conditions also give rise to unique strong solutions. One would then expect that naively applying an Euler type numerical scheme would lead to a similar divergence. As it turns out, not only is this the case but the interaction among the particles implies that if one particle diverges it can bring the other particles with it, thus the whole system diverges. We refer to this effect as *particle corruption*, which will be discussed in more detail in Chapter 9. Our goal is therefore to draw upon the work on standard SDEs to create a numerical scheme for MV-SDEs that is stable when one has non Lipschitz coefficients.

It should be noted here that taming is one solution to the problem of simulating SDEs with more general coefficients. It has been shown that the implicit (or backward) Euler scheme still converges for one sided Lipschitz functions, see [HMS02], [Szp10] and [MS13] for details on the algorithm. There are however, two problems extending implicit schemes here, firstly implicit schemes are far more costly than explicit schemes since they involve solving a fixed point equation at every step, see [HJK12]. Secondly, in order to simulate a MV-SDE (and hence a particle system) extending the implicit scheme to an N-particle system would then require N fixed point equations to solve at every step. This along with the already large computational growth in N suggests an implicit algorithm unfavourable in this setting. We shall return to this in Chapter 9.

2.2.2 Importance of Improved Simulation

As one may have realised, because we can introduce a particle system and write the discretisation error on that, the computational complexity of the step size is "the same" between SDEs and MV-SDEs. However, the same is not true as a function of N.

To see this let us consider an example and for simplicity let us ignore the discretisation error, hence the error is $O(1/\sqrt{N})$. For standard SDEs, due to the fact they are independent simulating N SDEs is N times more expensive than simulating a single SDE. Therefore, if one wants to decrease the error by one order of magnitude, then the simulation is 10^2 times the cost, that is the computational complexity grows like $O(N^2)$. For MV-SDEs all N particles depend on each other, therefore the cost of simulating N particles is N^2 times more expensive than simulating one. This implies that in order to decrease the error by one order of magnitude requires 10^4 times the cost, namely, the computational complexity grows like $O(N^4)$.

Clearly this growth makes large particle systems completely infeasible, hence one looks towards minimising the error with a fixed N. There are techniques known as *variance reduction* techniques for Monte Carlo which do this at the level of the statistical error. The goal of Chapter 10 is to use one such technique known as *importance sampling* to improve the performance. Importance sampling works by changing the measure under which the simulations are carried out, if one does this in such a way that "important" regions of the distribution are simulated from more often then one can dramatically reduce the variance. Of course in the MV-SDE setting the SDE depends on its own law (and hence measure) therefore one must be careful about how the measure change is performed. Although we focus purely on a reduction in the statistical error this method may provide an interesting approach to also reducing the propagation of chaos error, we shall discuss this further in Chapter 10.

2.2.3 Contributions

We make several contributions in these directions and provide answers to the question posed earlier. Firstly with regard to the simulation of MV-SDEs with super linear growth, in Chapter 9 we show additional difficulties in using basic Euler scheme, so-called *particle corruption* technique. We provide a pathwise propagation of chaos result in this setting, which had previously only been shown in the Lipschitz setting, see Proposition 9.1.2. We also propose and prove strong convergence of two algorithms capable of handling such MV-SDEs, an implicit and explicit, this is carried out in Theorems 9.1.9 and 9.1.3 respectively. We further obtain the standard $1/2$ rate of convergence w.r.t. the stepsize for the explicit scheme. Although we do not obtain a corresponding convergence rate for the implicit scheme, the numerical examples appear to suggest the explicit scheme is superior. Moreover, the proof of the implicit scheme is quite technical and relies on stopping time arguments. Due to the law dependency, stopping times are known to be an issue for MV-SDEs and hence we must be extremely careful applying such arguments.

The second part of our contribution is w.r.t. variance reduction for MV-SDEs. The idea of variance reduction techniques in Monte Carlo methods for standard SDEs is well studied, however, this is not the case for MV-SDEs. The added computational complexity of the particle makes more efficient techniques even more desirable. By efficient we mean, reduce the number of particles required to yield a given error. In Chapter 10 we propose two types of importance sampling algorithm (one is capable of handling super linear growth), see Section 10.1 for the algorithms. One of the first hurdles one must overcome here is, since the measure is an intrinsic part of the MV-SDE equation, but this explicit dependence is "lost" when taking a particle approximation, how does the measure change effect that? This to best of our knowledge has never properly been addressed and it doing it naively means that the particle system does not converge to the correct MV-SDE. To address this issue in Proposition 10.1.4 we present a propagation of chaos result under a measure change which shows the necessary adjustment one must make. For the importance sampling in Theorems 10.2.7 and 10.2.9 we use large deviation principles to select a measure change and show through numerical testing the variance reduction (and hence computational cost saving) one can achieve. Finally, this work is based solely on reducing Monte Carlo error, however, the theory may lend itself to reducing propagation of chaos error as well through a measure change designed to better sample from the law of the MV-SDE.

Chapter 3

Stochastic Methods for PDEs

The final area we study is the somewhat surprising connection between stochastic analysis and PDEs, namely, for certain classes of PDEs, one can represent the solution the PDE as the expected value of some function of an SDE. It is common to say a PDE whose solution can be written this way, has a *stochastic representation*. One useful result is that this gives access to Monte Carlo simulation which does not suffer the *curse of dimensionality*. There has been a vast amount of literature recently focusing on such representations, this is the main topic of Part IV.

3.1 Motivation and Background

It is difficult to overstate the importance of PDEs in mathematical modelling, they have found application in almost every area. For a complete review of the theory of PDEs, one can consult [Eva98] and references therein. For applications of PDEs in finance one can consult [GHL13] for example, where they derive and consider PDEs for many exotic options. While there have been many numerical methods considered to solve PDEs (see for example [Tre00]), we shall concentrate on so-called stochastic representations.

Many books involving SDEs contain a section detailing the connection between them and linear PDEs, often referred to as the *Feynman-Kac* formula. For this small exposition we follow the results in [PR16, Chapter 3.8], but one can obtain similar results in [Mao08], [KS12] and [Fri12] among others.

We start by defining the following SDE in $\mathbb{R}^d$, with fixed $x \in \mathbb{R}^d$ over the interval $[0,T]$ with $T > 0$ fixed,

$$\begin{cases} X_s^{t,x} = x, & 0 \le s \le t \\ X_s^{t,x} = x + \int_t^s b(s, X_s^{t,x})\mathrm{d}s + \int_t^s \sigma(s, X_s^{t,x})\mathrm{d}W_s, & t < s \le T, \end{cases}$$

where b, $\sigma : [0,T] \times \mathbb{R}^d \to \mathbb{R}^d$ are jointly continuous functions in t and x. The following conditions can be weakened (see [PR16, Chapter 3.7]), but to keep things simple let us assume the b and σ are Lipschitz in space and satisfy the usual time boundedness condition in time, namely $b(\cdot, 0) \in L^1$ and $\sigma(\cdot, 0) \in L^2$. Then $X^{t,x}$ has a unique strong solution (see [PR16, Theorem 3.17]).

Now let us also consider the coefficients, $c, f : [0,T] \times \mathbb{R}^d \to \mathbb{R}$ and $G : \mathbb{R}^d \to \mathbb{R}$, which we assume are continuous and there exists constants C, $p > 0$ such that,

$$|c(t,x)| \le C\,,\ |f(t,x)| + |G(x)| \le C(1 + |x|^p)\,,\ (t,x) \in [0,T] \times \mathbb{R}^d. \tag{3.1.1}$$

The basic idea of stochastic representations is to consider the following linear (backward) parabolic PDE, often referred to as the Cauchy problem,

$$\begin{cases} \partial_t u(t,x) + \mathcal{L}u(t,x) + c(t,x)u(t,x) + f(t,x) = 0 \quad (t,x) \in [0,T] \times \mathbb{R}^d, \\ u(T,x) = G(x), \quad x \in \mathbb{R}^d. \end{cases} \tag{3.1.2}$$

where $\mathcal{L}$ is the standard 2nd order differential operator, commonly referred to as an Itô generator or Dynkin operator, for functions $\varphi \in C^2(\mathbb{R}^d)$

$$(\mathcal{L}\varphi)(t,x) = \sum_{i=1}^{d} b_i(t,x)\frac{\partial \varphi}{\partial x_i}(x) + \frac{1}{2}\sum_{i,j=1}^{d}(\sigma\sigma^{\intercal})_{i,j}(t,x)\frac{\partial^2\varphi}{\partial x_i \partial x_j}(x)\,.$$

The Feynman-Kac formula for this problem is,

$$u(t,x) := \mathbb{E}\left[G(X_T^{t,x})e^{\int_t^T c(s,X_s^{t,x})\mathrm{d}s} + \int_t^T f(s,X_s^{t,x})e^{\int_t^s c(r,X_r^{t,x})\mathrm{d}r}\mathrm{d}s\right]. \tag{3.1.3}$$

The goal then is to connect (3.1.3) with the solution to (3.1.2). We have the following result [PR16, Proposition 3.41].

Proposition 3.1.1. *Let $u \in C^{1,2}([0,T]\times\mathbb{R}^d)$ be a solution to* (3.1.2) *such that there exists a M, $q>0$ with,*

$$|u(t,x)| \le M(1+|x|^q)\,, \quad \forall (t,x)\in[0,T]\times\mathbb{R}^d.$$

Moreover, if the growth conditions (3.1.1) *are satisfied, then $u(t,x)$ satisfies* (3.1.3) *(The Feynman-Kac formula).*

We give an idea of the proof, for $d=1$, firstly consider the derivative on the following product,

$$\begin{aligned}
\mathrm{d}\left(u(s,X_s^{t,x})e^{\int_t^s c(r,X_r^{t,x})\mathrm{d}r}\right) &= \mathrm{d}\left(u(s,X_s^{t,x})\right)e^{\int_t^s c(r,X_r^{t,x})\mathrm{d}r} + u(s,X_s^{t,x})\mathrm{d}\left(e^{\int_t^s c(r,X_r^{t,x})\mathrm{d}r}\right)\\
&= e^{\int_t^s c(r,X_r^{t,x})\mathrm{d}r}\Big[\big(\partial_t u(s,X_s^{t,x}) + b(s,X_s^{t,x})\partial_x u(s,X_s^{t,x}) + \frac{\sigma(s,X_s^{t,x})^2}{2}\partial_{xx}u(s,X_s^{t,x})\big)\mathrm{d}s\\
&\quad + \sigma(s,X_s^{t,x})\partial_x u(s,X_s^{t,x})\mathrm{d}W_s\Big] + u(s,X_s^{t,x})e^{\int_t^s c(r,X_r^{t,x})\mathrm{d}r}c(s,X_s^{t,x})\mathrm{d}s\,,
\end{aligned}$$

where we applied Itô's formula to u. The idea is to write this in integral form and take expectations,

$$\begin{aligned}
\mathbb{E}\left[u(T,X_T^{t,x})e^{\int_t^T c(r,X_r^{t,x})\mathrm{d}r}\right] &= u(t,x) + \mathbb{E}\Big[\int_t^T e^{\int_t^s c(r,X_r^{t,x})\mathrm{d}r}\Big[\partial_t u(s,X_s^{t,x}) + b(s,X_s^{t,x})\partial_x u(s,X_s^{t,x})\\
&\quad + \frac{\sigma(s,X_s^{t,x})^2}{2}\partial_{xx}u(s,X_s^{t,x}) + u(s,X_s^{t,x})c(s,X_s^{t,x})\Big]\mathrm{d}s\Big]\\
&= u(t,x) + \mathbb{E}\Big[\int_t^T e^{\int_t^s c(r,X_r^{t,x})\mathrm{d}r}(-f(s,X_s^{t,x}))\mathrm{d}s\Big]\,,
\end{aligned}$$

where we have used the expectation to remove the stochastic integral (although technically one needs to show enough integrability) and used that u solves the PDE to obtain the final line. Noticing that $u(T,X_T^{t,x}) = G(X_T^{t,x})$ and rearranging for $u(t,x)$ one obtains (3.1.3).

There are of course some quite strong assumptions in here such as u being a unique classical solution, but as it turns out, these assumptions can be lifted and we can connect (3.1.3) with (3.1.2), provided one is willing to also accept a weaker form of solution, a so-called *viscosity solution* of the PDE, which we will return to in Chapter 11. This then leads to the far more general result [PR16, Theorem 3.42],

Theorem 3.1.2 (Feynman-Kac's Formula). *Let the coefficients in the PDE* (3.1.2) *be continuous and satisfy the growth conditions* (3.1.1). *Then the function u defined in* (3.1.3) *is a continuous function of $(t,x)\in[0,T]\times\mathbb{R}^d$ which grows at most polynomially at infinity and is the unique viscosity solution of* (3.1.2) *satisfying,*

$$\lim_{|x|\to\infty}|u(t,x)|e^{-\delta[\log(|x|)]^2} = 0\,,$$

uniformly for $t\in[0,T]$ and some $\delta>0$.

This result is one of the reasons stochastic representations are useful, even when the PDE does not have a classical solution, (3.1.3) still has some meaningful connection and hence is the basis for the majority of the theoretical work in the area. We do not give details of the proof here but the interested reader can consult [PR16] for further discussion.

Remark 3.1.3 (Connections with different types of PDE). *The above results apply to backward PDEs i.e. PDEs with a terminal condition. As it turn out one can construct analogous results for PDEs with initial conditions provided that the coefficients in the SDE, as well as that in the PDE do not depend on time. See* [PR16, Chapter 3.8] *for details and results. One can also consult this text for adding in Dirichlet Boundary conditions and elliptic PDEs.*

We now move on to generalisations of the Feynman-Kac formula and show the limitation using standard SDEs. The above stochastic representation was useful since it allowed us to represent the solution of the PDE purely in terms of its coefficients. Let us consider the more general nonlinear PDE (often referred to as a semi-linear PDE),

$$\begin{cases} \partial_t u + \mathcal{L}u + f(t,x,u(t,x),\nabla_x u(t,x)\sigma(t,x)) = 0\,, & (t,x) \in [0,T) \times \mathbb{R}^d\,, \\ u(T,\cdot) = G(\cdot)\,, & x \in \mathbb{R}^d\,. \end{cases} \tag{3.1.4}$$

Our goal is to derive the stochastic representation for this PDE. For the sake of argument let us assume that there is a unique $C^{1,2}$ solution and nice conditions on all coefficients. Following a similar argument to above we would obtain the following stochastic representation,

$$u(t,x) := \mathbb{E}\left[G(X_T^{t,x}) + \int_t^T f\big(s, X_s^{t,x}, u(s,X_s^{t,x}), \nabla_x u(s,X_s^{t,x})\sigma(s,X_s^{t,x})\big)\mathrm{d}s \right]\,. \tag{3.1.5}$$

This is a problem since now the stochastic representation depends on the solution itself. That is (3.1.3) depends only on known functions c, f and G, while (3.1.5) depends on the unknown u. This is clearly not desirable and shows the limitation of using SDEs as stochastic representations.

The breakthrough to this more general framework was given by Backward Stochastic Differential Equations (BSDEs) and Forward BSDES (FBSDEs), see [PR16], [PP92], [Car08, Chapter 8], [EKPQ97], [CM10] among others for results on (F)BSDEs generally and their application to PDEs. Although BSDEs give a solution, from a numerical standpoint they can be difficult to work with. A different approach to deal with nonlinear PDEs (although under more restrictive assumptions) is branching diffusions, originally developed by [Sko64], [Wat65], [McK75], however more recently [RRM10], [HL12], [HLTT14], [HLOT$^+$16] have extended the scope. Both of these techniques require a good deal of machinery in order to properly convey the idea and results, we therefore leave this here and give a detailed explanation of both FBSDEs and branching diffusions in Chapters 11 and 12.

3.1.1 Why Stochastic Representations are numerically useful

As mentioned above there are some theoretical merits from using stochastic representations of PDEs, however, there are large gains to be made numerically from such representations. This motivation follows the author's joint work with Bernal and dos Reis [BdRS17].

As mentioned PDEs are ubiquitous in modelling, appearing in image analysis and processing, inverse problems, shape analysis and optimization, filtering, data assimilation and optimal control. They are used in Math–Biology to model population dynamics with competition or growth of tumours; or to model complex dynamics of movement of persons in crowds or to model (ir)rational decisions of players in games and they feature in many complex problems in Mathematical Finance. Underpinning all these applications is the necessity of solving numerically such equations either in bounded or unbounded domains in potentially high dimensions. Firstly one should note that Monte Carlo does not suffer from the curse of dimensionality as deterministic based solvers do. This is shown in [HJK16], where the authors use stochastic techniques to solve a 100 dimensional semilinear parabolic PDE. Solving these types of problems in almost unthinkable with traditional PDE solvers. However, where one is able to make real computational savings is through parallelising the PDE solver which we shall now explain.

Deterministic Domain Decomposition. Let us start by attempting to parallelise a classic (deterministic) based solver. The standard example of a Boundary Value Problem (BVP) is Laplace's equation with Dirichlet Boundary Conditions (BCs):

$$\Delta u(x) = 0 \text{ if } x \in D \subset \mathbb{R}^d, \qquad u(x) = g(x) \text{ if } x \in \partial D. \tag{3.1.6}$$

The large data sets involved in realistic applications nearly always imply that the discretisation of a BVP such as (3.1.6) leads to algebraic systems of equations that can only be solved on a parallel computer with a large number (say $p >> 1$) of processors. Not only does parallelisation require multiple processors but also parallel algorithms. The classical Schwarz's alternating method was the first and remains the paradigm of such algorithms which we refer to as "Deterministic Domain Decomposition" (DDD) [SBG04]. While state-of-the-art DDD algorithms outperform Schwarz's alternating method in every respect, the latter nonetheless serves to illustrate the crucial difficulty they all face. The idea of Schwarz's algorithm is to divide D into a set of p overlapping subdomains, and have processor $j = 1, \cdots, p$ solve the restriction of the PDE to the subdomain, D_j see Fig. 3.1.

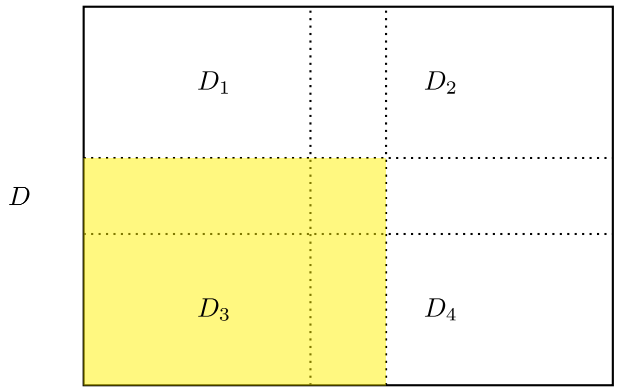


Figure 3.1: Domain decomposition on an arbitrary domain D, split into four overlapping subdomains D_i as required for Schwarz's alternating method. The subdomain D_3 is highlighted.

Since the solution is not known in the first place, the BCs on the fictitious interfaces of D_j are also unknown, therefore, an initial guess has to be made in order to give processor j a well-posed (yet incorrect) problem. The BCs along the fictitious interfaces of D_j are then updated from the solution of the surrounding subdomains in an iterative way until some convergence criteria is met.

Inter-processor communication and the scalability limit of DDD. Since the inter-processor communication involved in DDD's updating procedure is intrinsically sequential, it sets a limit to the scalability of the algorithm by virtue of the well known Amdahl's law.

A simple illustrative example is as follows. A fully scalable algorithm would take half the time (say $T/2$) to run if the number of processors was doubled. If there is a fraction $\nu < 1$ of the algorithm which is sequential, then the completion time will not drop below $T\nu$, regardless of how many processors are added. For instance, in Schwarz's method, if 5% (i.e. $\nu = .05$) of the execution time of one given processor is lost by waiting for the artificial BCs to be ready, then the execution time could be shortened by at most a factor of 20 (with infinitely many processors; a factor of about 19 would already take around 360 processors[*]. In other words, Schwarz's alternating algorithm, or *any* DDD algorithm for that matter, cannot exploit the full capabilities of a parallel computer due to the idle time wasted in (essential) communication.

We emphasise this point further by borrowing an example from David Keyes[†]. The Gordon Bell prizes are annually awarded to numerical schemes which achieve a breakthrough in performance when solving a realistic problem. In 1999, one such problem was the simulation of

[*] The time taken can be theoretically calculated by the amount of extra overlap caused by the processors divided by the number of processors, namely $T(1 + (p-1)\nu)/p$.

[†] See http://www.mcs.anl.gov/research/projects/petsc-fun3d/Talks/bellTalk.ppt

the compressible Navier-Stokes equations around the wing of an airplane. With 128 processors, the winning code took 43 minutes to solve the task. On the other hand, with 3072 processors it took it 2.5 minutes instead of 1.79 as would have been the case with a fully parallelisable algorithm. The remaining 28% of computer time were lost to interprocessor communication. At this point, adding more processors would have led to a faster loss of scalability.

The Probabilistic Domain Decomposition (PDD) method. A conceptual breakthrough was achieved by Acebrón *et al.* with the PDD method (or rather, the PDD framework) [ABLS05], based on a previous, unpublished idea by Renato Spigler. PDD is the only domain decomposition method potentially free of communication, and thus potentially fully scalable. It does so by splitting the simulation in two separate stages, the first stage recasts the BVP into a stochastic formulation (via the Feynman-Kac formula) which allows to compute the solution of the PDE at certain specific points in time/space. Thus, we can compute the "true" solution values of the DDD's fictitious interfaces for the D_j's. Therefore, the fictitious boundaries that were previously unknown are now known and hence do not need to be communicated!

Consequently, the subdomains are now completely independent of each other, the second stage then involves solving for the solution over the subdomains in a full parallel way. PDD calculations will be affected by two independent sources of numerical error: the subdomain solver and the statistical error of Monte Carlo simulations.

Results for the speed up one can obtain and further results on the algorithms can be found in works such as [ARRS09], [ARRS10], [BA16], [BdRS17] among others.

Key Questions.

There are several questions we wish to address here.

- How can one solve more general PDEs than linear second order via so-called branching diffusions? Related to this, how can such representations be used to construct unbiased estimators?
- How can we then use the branching diffusion machinery to consider a representation for a first order PDE?
- Can this representation be used/adapted to solve more general PDEs, such as degenerate second order PDEs?

To answer these questions requires us to introduce additional theory, we therefore postpone the discussion until Part IV.

3.1.2 Contributions

The main purpose of Chapter 12 is to give an overview of branching diffusions. This is mainly to help the reader unfamiliar with the topic since the current literature is limited to a handful of papers and is technical.

We have discussed why stochastic representations are useful, but as one may observe from our examples, we always require a second order term in space. In Chapter 13 we shall remove this requirement and construct a stochastic representation for a first order PDE which has finite variance and is unbiased, see Theorem 13.2.7. As far as we are aware such a result is the first of its kind and although it has been carried out on a simple first order PDE there is some numerical evidence to suggest this method can be extended to the nonlinear setting. Such a breakthrough has much wider implications since such a method would be capable of tackling nonlinear first order PDEs in high dimensions, or indeed degenerate second order semilinear PDEs which are currently out of reach from the current theory.

Chapter 4

Notation

Due to the range of topics considered in some instances we use the same symbol to denote something different in different parts. Although there are several notations that we keep consistent throughout which we list here.

Notation

We denote the set of natural numbers (without zero) by $\mathbb{N}$ and unless otherwise stated d, n, m, l, N, M are positive integers i.e. $\in \mathbb{N}$. Further, we will work with $\mathbb{R}^d$, the d-dimensional Euclidean space of real numbers, and for $a = (a_1, \cdots, a_d) \in \mathbb{R}^d$ and $b = (b_1, \cdots, b_d) \in \mathbb{R}^d$ we denote by $|a|^2 = \sum_{i=1}^d a_i^2$ the usual Euclidean distance on $\mathbb{R}^d$ and by $\langle a, b\rangle = \sum_{i=1}^d a_i b_i$ the usual scalar product. To denote m-by-d real matrices, we write $\mathbb{R}^{m\times d}$, for $A \in \mathbb{R}^{m\times d}$, we denote its transpose by $A^\intercal$. Let $a, b \in \mathbb{R}$, we use the standard probability notation for of $a \wedge b$ to denote $\min(a, b)$ and $a \vee b$ to denote $\max(a, b)$.

As is standard in analysis, we use C to denote a generic positive constant that can change from line to line. Crucially C will only ever depend on "known" parameters e.g. Lipschitz constants, T etc. In the case C depends on a changeable parameter, for example x, we write $C(x)$.

We denote by $\mathcal{B}$ the Borel σ-algebra, and when it is required to be explicit on which space we use $\mathcal{B}(\mathbb{R}^d)$ for example. For two measurable space $(X, \mathcal{E})$ and $(Y, \mathcal{G})$ we denote by $\otimes$ the standard product σ-algebra of the measurable spaces. Namely, $\mathcal{E} \otimes \mathcal{G}$ is the σ-algebra generated from the Cartesian product of sets in $\mathcal{E}$ and $\mathcal{G}$. For a given (measurable) set A, we denote by $\mathbb{1}_A$ the indicator of that set.

The parameter t is always taken as time and we work on a finite interval, i.e. we set some $0 < T < \infty$ and take $t \in [0, T]$. We denote by $(\Omega, \mathcal{F}, \mathbb{P})$ a probability space and where appropriate the filtered probability space, $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \mathbb{P})$. In the case where we are dealing with Brownian motion Parts III and Part IV, $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \mathbb{P})$ is taken to satisfy "the usual conditions", where $\mathcal{F}_t$ is the augmented filtration of a standard multidimensional Brownian motion W on the time interval $[0, T]$.

$\mathbb{P}$ and $\mathbb{Q}$ are probability measures and $\mathbb{E}$ is the expectation (under $\mathbb{P}$ unless stated otherwise). When we need to be explicit we write $\mathbb{E}_\mathbb{P}$ or $\mathbb{E}_\mathbb{Q}$ to denote the expectation under this measure. Similarly, W is often taken as a Brownian motion under $\mathbb{P}$, when we need to be explicit we write $W^\mathbb{P}$ or $W^\mathbb{Q}$ to denote which measure it is a Brownian motion under.

Function Spaces

We use the following notation for spaces, which are standard in stochastic analysis literature. We define $\mathbb{S}^p$ for $p \geq 1$, as the space of $\mathbb{R}^d$-valued, $\mathcal{F}_\cdot$-adapted processes Z, that satisfy, $\mathbb{E}[\sup_{0\leq t\leq T} |Z(t)|^p]^{1/p} < \infty$. Similarly, $L_t^p(\mathbb{R}^d)$, defines the space of $\mathbb{R}^d$-valued, $\mathcal{F}_t$-measurable random variables X, that satisfy, $\mathbb{E}[|X|^p]^{1/p} < \infty$.

We denote by $C(\mathbb{R}^d)$ as the set of continuous functions on $\mathbb{R}^d$, we further denote by $C_b^k(\mathbb{R}^d)$ all k times differentiable real valued functions defined on $\mathbb{R}^d$, with bounded partial derivatives up to order k. Similarly, we denote by $C^{1,2}([0, T] \times \mathbb{R}^d)$ all real valued functions defined on the product space $[0,T] \times \mathbb{R}^d$ which are continuously differentiable in the first component and twice continuous differentiable in the second. We use $C_0([0,T])$ to denote all continuous real valued functions on the interval $[0,T]$, with value zero at (time) zero. When it is not clear we use ";" to denote what space is being mapped from and to, for example $C(X;Y)$ is the continuous functions mapping X to Y.

Finally we denote by $\mathbb{H}_T^d$ the Cameron-Martin space of absolutely continuous functions with square integrable derivative over the interval $[0,T]$ with values in $\mathbb{R}^d$, i.e. (if $d=1$ we just write $\mathbb{H}_T = \mathbb{H}_T^1$)

$$\mathbb{H}_T^d = \left\{ h : [0,T] \mapsto \mathbb{R}^d : \ h_0 = 0\,,\ h_\cdot = \int_0^\cdot \dot{h}_t dt\,,\ \int_0^T |\dot{h}_t|^2 \, \mathrm{d}t < \infty \ \text{ i.e. } \ \dot{h}_t \in L_t^2(\mathbb{R}^d) \right\}.$$

Part II

Credit Risk Modelling

Chapter 5

Preliminaries

This part is based on the author's joint work with dos Reis [dRS17] and dos Reis and Pfeuffer [PRS18].

The goal in this part is to develop techniques to help improve the estimation of probabilities that are crucial in credit risk modelling, most importantly probability of default. Initially, in Chapter 6 we work under the Markov assumption, that is we treat every company in the same rating as having the same risk. This is also the best modelling one can come up with when individual transitions are not known. Following that, in Chapter 7 we consider the problem when one has access to the full list of company transitions and therefore can remove the Markov assumption. This makes the theory more complex, however, we are able to better capture the data using these techniques. In both cases our goal will be to calibrate a model and we opt for statistical methods to do this and in particular likelihood inference.

5.1 Markov Chains

The theory of Markov chains (processes) is apparent in many areas of mathematics and science, although here we will only provide minimal details of the concepts and results needed for TPMs. We will focus on continuous time Markov chains (CTMCs) on a discrete state space. For more details and proofs of these one should consult one of the many books on the subject, for example [Nor98].

We work with a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and consider a stochastic process $X = \{X(t) | \, 0 \leq t\}$ defined on a finite discrete state space S, namely we can define $S := \{1, \dots, h\}$ for $h \in \mathbb{N}$. Hence for any $t \geq 0$, $X(t) : \Omega \to S$, we say that such a process is a CTMC if the following property holds for all $t, s \geq 0$, and $i, j \in S$,

$$\mathbb{P}\big[X(s+t) = j | X(s) = i, \, \{X(u) : 0 \leq u \leq s\}\big] = \mathbb{P}\big[X(s+t) = j | X(s) = i\big] \,. \tag{5.1.1}$$

Further, we say the process is a time-homogeneous CTMC if the following also holds,

$$\mathbb{P}\big[X(s+t) = j | X(s) = i\big] = \mathbb{P}\big[X(t) = j | X(0) = i\big] \,. \tag{5.1.2}$$

One key concept in CTMCs is the notion of a generator matrix, which we will come onto very shortly, but first let us consider the following, for any matrix Q the series,

$$\sum_{k=0}^{\infty} \frac{Q^k}{k!} \,, \tag{5.1.3}$$

converges component-wise and we denote the limit of such a series as e^Q (naturally). Moreover, suppose for some matrix P, that we can find a matrix Q, such that $e^Q = P$. Then, for all $t \geq 0$, we obtain the relation $e^{tQ} = P^t$. This leads to the following theorem (note at present there are no restrictions on P and Q).

Theorem 5.1.1. *Let Q be a matrix on S. Set, $P(t) = e^{tQ}$. Then $(P(t) : t \geq 0)$ has the following* *properties:*

1. $P(s+t) = P(s)P(t)$ *for all* s, t *(semigroup property).*
2. $(P(t) : t \geq 0)$ *is the unique solution to the forward equation*

$$\frac{d}{dt}P(t) = P(t)Q\,, \quad P(0) = Id\,. \tag{5.1.4}$$

3. $(P(t) : t \geq 0)$ *is the unique solution to the backward equation*

$$\frac{d}{dt}P(t) = QP(t)\,, \quad P(0) = Id\,. \tag{5.1.5}$$

4. *for* $k = 0, 1, 2, \ldots$*, we have,*

$$\left(\frac{d}{dt}\right)^k \bigg|_{t=0} P(t) = Q^k\,. \tag{5.1.6}$$

The beauty of this theorem is that there are very little assumptions placed on P and Q. However, let us define the following [Nor98, pg 69].

Definition 5.1.2. *Consider some set S and a matrix Q, Q is referred to as a generator (or Q-matrix) on this set if the following conditions hold:*

- $0 \leq -q_{ii} < \infty$ *for all* $i \in S$.
- $q_{ij} > 0$ *for all* $i, j \in S$ *such that* $i \neq j$.
- $\sum_{j \in S} q_{ij} = 0$ *for all* $i \in S$.

This implies $q_{ii} = -\sum_{j \neq i} q_{ij}$*, i.e. rows of Q all sum to zero. Further, it is common to denote* $q_i = \sum_{j \neq i} q_{ij}$ *(as an aside q_i is taken to be the rate of leaving state i).*

This leads us onto the extremely useful theorem [Nor98, pg 63].

Theorem 5.1.3. *A matrix Q on a finite set S is a generator iff $P(t) = e^{tQ}$ is a stochastic matrix for all $t \geq 0$.*

Note that Theorem 5.1.3 implies if Q is a generator, then $P := e^Q$ is a stochastic matrix. However, given a stochastic matrix P, there may not exist a generator Q that $e^Q = P$. This is what gives rise to Theorem 1.2.1 and will be discussed further in Chapter 6.

The Q matrix also implies how one can simulate a Markov chain, we observe that state i has intensity q_i, that is, if $X = i$, then the probability of X moving follows an exponential distribution with mean $1/q_i$. Moreover, when X transitions to another state the probability of jumping to state j is q_{ij}/q_i.

Likelihood of CTMC

Let $(X(t))_{t \geq 0}$ denote a CTMC over the finite state space $\{1, \ldots, h\}$ with a generator Q. Associated to $X(t)$ is, for i, j in the state space, $K_{ij}(t)$ the number of jumps from i to j in the interval $[0, t]$ and by $S_i(t)$ the holding time of state i in the interval $[0, t]$.

The likelihood of a continuous time fully observed Markov chain with generator Q is given by the following expression (see [KS97, Chapter 3.4]),

$$L_t(Q) = \exp\Big(\sum_{i=1}^{h} \Big[\sum_{j \neq i} \log(q_{ij}) K_{ij}(t) - S_i(t) \sum_{j \neq i} q_{ij}\Big]\Big) = \prod_{i=1}^{h} \Big(\prod_{j \neq i} q_{ij}^{K_{ij}(t)}\Big) \exp\big(-q_i S_i(t)\big)\,. \tag{5.1.7}$$

One can see from this expression why we have a "missing data" problem. In the case of the discretely observed TPMs, we not observe K and S, hence the likelihood cannot be fully evaluated and by extension cannot simply be maximised.

5.2 Overview of Markov Chain Monte Carlo algorithm

For details on the Markov Chain Monte Carlo (MCMC) theory we refer the reader to [GRS96]. Algorithms for implementing MCMC to estimate a generator, from discrete observations are discussed in [BS05] and [BS09]. MCMC differs from the Expectation Maximisation (EM) in the sense that EM estimates the set of parameters which maximises the likelihood function (see Section 6.1), while MCMC samples from the posterior distribution. Namely, given some data D, the posterior distribution of parameters θ is $\pi(\theta|D)$, which by Bayes' theorem is,

$$\pi(\theta|D) = \frac{\pi(D|\theta)\pi(\theta)}{\int \pi(D|\theta)\pi(\theta)d\theta} ,$$

with $\pi(D|\theta)$ denoting the likelihood and $\pi(\theta)$ the prior distribution. MCMC obtains the best guess of θ by sampling from $\pi(\theta|D)$ and taking the Monte Carlo approximation of the expected value. The reason the expectation is our best guess is due to the fact we use both the data (likelihood) but also our experience on what the outcome should approximately be (the prior). Although the prior can be extremely useful in stopping 'bad' answers it is also a criticism of MCMC due to so-called prior sensitivity.

5.2.1 MCMC with missing data

As alluded to in the previous section, the problem one faces here is a missing data problem. One algorithm often used in this situation is the Expectation Maximisation (EM) algorithm which we discuss further in Section 6.1. However here we focus on the MCMC approach.

Remark 5.2.1. *Here we purely discuss MCMC to sample from the posterior, this is the standard approach. MCMC algorithms which approximate the maximum likelihood in the presence of missing data do exist, but are more useful when for example one cannot explicitly write the E step in the EM algorithm (see* [GC93]*).*

Similar to the case of the EM algorithm the problem faced here is missing data. Namely we wish to consider the so-called posterior distribution of the generator matrix Q, which we denote by $\pi(Q|D)$ (although it is common to suppress the data and only write $\pi(Q)$). The difficulty is, in its current state this is an extremely hard distribution to evaluate since we do not have a good handle on the likelihood, so we augment with an auxiliary variable X (see [GRS96, p.105] and [BG93]). In general X need not require an interpretation, although here it will correspond to the full Markov chain. In order to generate realisations of $\pi(Q|D)$, we specify the conditional distribution $\pi(X|Q, D)$ which provides the joint distribution $\pi(Q, X|D) = \pi(Q|D)\pi(X|Q, D)$ and therefore the marginal distribution of Q is $\pi(Q|D)$. One can then sample from the marginal distribution by using any sampling method that preserves the joint distribution $\pi(Q, X|D)$ (and by extension $\pi(Q|D)$), such as Gibbs or Metropolis Hastings.

The method used in [BS05] and [BS09] is the data augmentation algorithm from [TW87] (see also [LR02, p.200]). We specify the prior distribution $\pi(Q)$ and take a realisation from this distribution, $Q^{(0)}$, we then construct a sequence $\{Q^{(k)}, X^{(k)}\}$, for $k = 1, \dots, M$ by:

1. Draw, $X^{(k)} \sim \pi(X|Q^{(k-1)}, D)$.
2. Draw, $Q^{(k)} \sim \pi(Q|X^{(k)}, D) = \pi(Q|X^{(k)})$ (since $X^{(k)}$ is richer than D).
3. Save $\{Q^{(k)}, X^{(k)}\}$ and take $k = k + 1$.

Under mild conditions (see [GRS96, Chapter 4]), after some burn-in n, the sequence $\{Q^{(k)}, X^{(k)}\}$ for $k \geq n$ has the same distribution as $\pi(Q, X|D)$. Moreover, the marginals also have the correct distribution, namely, $\{Q^{(k)}\} \sim \pi(Q|D)$ for $k \geq n$. Therefore we estimate the generator matrix by, $\frac{1}{M-n+1}\sum_{k=n}^{M} Q^{(k)}$.

For the choice of prior, $\pi(Q)$, [BS05] suggest a prior from the gamma distribution with shape α_{ij} and scale $1/\beta_i$. Hence, $q_{ij} \sim \Gamma(\alpha_{ij}, 1/\beta_i)$, where $\alpha_{ij}, \beta_i \geq 0, \forall\, i \neq j \in \{1, \dots, h\}$. With this choice, the prior is a conjugate prior. Although this prior has some drawbacks, we note, by assuming the prior to follow a Gamma distribution we effectively bound the parameter space, therefore there is no need to make the space compact. Noting that the posterior distribution of X is equivalent to the likelihood i.e. $\pi(X|Q) = L_t(X; Q)$, one has

$$\pi(Q|X, D) = \pi(Q|X) = \frac{\pi(Q, X)}{\pi(X)} \propto L_t(X; Q)\pi(Q)\,.$$

From the likelihood of a CTMC, (5.1.7) and the assumption on the prior we infer that,

$$\begin{aligned} L_t(X; Q)\pi(Q) &\propto \prod_{i=1}^{h}\prod_{j\neq i} q_{ij}^{K_{ij}(t)} e^{-S_i(t)q_{ij}} \prod_{i=1}^{h}\prod_{j\neq i} q_{ij}^{\alpha_{ij}-1} e^{-\beta_i q_{ij}} \\ &= \prod_{i=1}^{h}\prod_{j\neq i} q_{ij}^{K_{ij}(t)+\alpha_{ij}-1} e^{-(S_i(t)+\beta_i)q_{ij}}\,. \end{aligned}$$

That is, X gives us both K_{ij} and S_i. We do not have equality here since there is no normalisation term, however, note that the resulting posterior is in the same form as the prior. Hence we generate q_{ij} with $i \neq j$ from the distribution $\Gamma\big(K_{ij}(t) + \alpha_{ij}, 1/(S_i(t) + \beta_i)\big)$ (since each q_{ij} is taken as independent).

5.2.2 Single Component Metropolis Hastings

The previous description deals with the MCMC algorithm in the Markov missing data setting, which we use in Chapter 6. However, in Chapter 7 we have access to the full data set (that is we know K and S for each individual company) and therefore we do not require the data augmentation step. That being said, because we have access to individual company transitions we look to drop the Markov assumption and use a more complex model, which in turn implies we do not have the conjugate priors as described above. This forces us to then use a more general sampling method, the so-called Metropolis Hastings (MH) algorithm, originally developed in this context by [Has70]. In our case we have multiple parameters to estimate, hence we look to use a variant of the MH, the so-called single component MH, described in [GRS96, Section 1.4].

The single component MH algorithm works by preforming n updating steps for each time interval. Let us denote by* $X_{t,i}$ the state of component i at time t. The next iteration for step i, i.e. the $t + 1$ iteration is then carried out as a standard MH algorithm, we generate a candidate point Y_i from the distribution, $\psi_i(Y_i|X_{t,i}, X_{t,-i})$, where $X_{t,-i}$ denotes the most recent step in the $t + 1$ iteration, namely,

$$X_{t,-i} := \{X_{t+1,1}, \dots X_{t+1,i-1}, X_{t,i+1}, \dots, X_{t,n}\}\,, \tag{5.2.1}$$

that is, components $1, \dots, i - 1$ have already been updated. Therefore, the i^{th} proposal distribution only generates a candidate point for the i^{th} component of X. However, it may do this with dependence on the other components of X. The acceptance of any such component happens with probability $\alpha(X_{-i}, X_i, Y_i)$, which is defined as,

$$\alpha(X_{-i}, X_i, Y_i) = \min\left(1, \frac{\pi(Y_i|X_{-i})\psi_i(X_i|Y_i, X_{-i})}{\pi(X_i|X_{-i})\psi_i(Y_i|X_i, X_{-i})}\right)\,,$$

where $\pi(X_i|X_{-i})$ is the *full conditional* distribution of X_i under $\pi(\cdot)$ (which we define below). As one would expect, if Y_i is accepted or rejected we only update the i^{th} component accordingly, the other components do not change.

Remark 5.2.2 (More Advanced Algorithms). *Here we presented classical MCMC, which works well in many situations but can be slow. There is vast amounts of work on extending these sorts of algorithms with so-called Hamiltonian (or Hybrid) Monte Carlo (HMC)* [Nea11]*, Metropolis Adjusted Langevin Algorithm (MALA)* [RT96]*, Sequential Monte Carlo (see* [GDF01]*) and Particle MCMC (see* [DM13]*). For our purposes, however, MCMC will suffice.*

[*] Note that here X is still a Markov chain, except now it is defined on a general state space rather than the discrete one we were working with in Section 5.1. Also its purpose here is to obtain model parameter distributions rather than firm transitions.

5.3 Risk Measures

As so much of finance relies so heavily on the idea of risk, let us describe the concept of a risk measure. Rather simply a risk measure is a way to express loss in terms of risk. One can consult [Lüt08, Chapter 2] or [MFE15] for more details.

The most well known risk measure is *value at risk* (VaR), which mainly stems from the fact it is easy to understand. It can be defined in the following way for portfolio P with (random) loss L within α percent,

$$\mathrm{VaR}_\alpha(P) = \inf_{x\in\mathbb{R}} \{\mathbb{P}[L \le x] \ge \alpha\} \,. \tag{5.3.1}$$

That is, if we have some probability distribution of possible loss, VaR_α is the smallest loss such that with α percent the loss observed does not exceed VaR_α.

Although, VaR is intuitive one can ask does such a measure behave as expected? Related to this [ADEH99] was the first to lay down a solid mathematical framework for risk measures. With the main contribution being the idea of *coherent* risk measures. Although we won't go into the details about coherent risk measures, it is essentially a set of properties one expects a "good" risk measure to satisfy. One such property is diversification and as it turns out, VaR does not satisfy this property.

Due to this, the regulation is changing to so-called *Expected Shortfall* (ES) which is a coherent measure. ES is in fact closely related to VaR, except instead of one using a fixed level α, one averages over all confidence levels $\beta \ge \alpha$ for a fixed $\alpha \in (0,1)$, that is,

$$\mathrm{ES}_\alpha(P) = \frac{1}{1-\alpha}\int_\alpha^1 \mathrm{VaR}_\beta(P)\mathrm{d}\beta \,.$$

It is of course clear that $\mathrm{ES}_\alpha \ge \mathrm{VaR}_\alpha$ and by conditioning over all remaining confidence levels, ES better deals with "tail risk" and also does not suffer the discontinuity of VaR. That is (especially for small portfolios), one can have $\mathrm{VaR}_{\alpha+\epsilon} \gg \mathrm{VaR}_\alpha$ for small ϵ (see [GLS00]).

5.4 Point Processes

In the Markov chain setting we had a constant intensity, that is, q_i denoted the "rate" at which one exits state i. This is of course part of the Markov framework since the intensity is completely determined by the current state. Therefore, to include non-Markov effects into the model, namely rating momentum we must look to generalise this. Of course the Markov framework appears very suitable for this problem and hence we do not want to change too many aspects and this leads one (rather naturally) to *point processes*. Point processes is a vast topic in itself and we point the interested read to [DVJ03] and [DVJ07] for a full overview on the theory and applications.

Before getting into too much detail let us pick up from Chapter 1 and discuss how we look to embed history dependence into the model. Recall that we were interested in Hawkes processes (a specific type of *self-exciting* point process) which have intensities of the form

$$\lambda_t = \mu + \int_0^t \phi(t-s)\mathrm{d}N_s \,. \tag{5.4.1}$$

By setting $\phi = 0$ then we have constant intensity and this is equivalent to the Markov setting. However, ϕ allows us to vary the intensity with previous events, this is a key feature we need for momentum. As also described in Chapter 1, at this point the Hawkes process is just a counting process (it is a generalisation of a Poisson process), to make a process that takes values on some state space we consider a so-called *marked point process*, see [DVJ03, Section 6.4]. Marked point processes (MPPs), are essentially point processes on a product space $\mathcal{T} \times \mathcal{K}$, that is, we return a set of values, $\{t_k, \kappa_k\}$ for $k = 1, 2, \dots,$ where we typically think of t_k as the event time of the point process (with intensity λ) and κ_k of the "mark" associated to the event. Although these notions are what we shall use, in general t_k could be a multidimensional object e.g. include spatial dependence. In our particular case we shall have $\kappa_k \in \{1, \dots, h\}$ namely, it will denote the rating and this ensures our marked point process is well defined and is sometimes referred to as a multivariate point process in this case.

The likelihood of MPP, is given in [DVJ03, p.251],

$$L = \prod_{i=1}^{N_g(T)} \lambda_g^*(t_i) f^*(\kappa_i | t_i) e^{-\int_0^T \lambda_g^*(u) \mathrm{d}u}\,, \tag{5.4.2}$$

where we have the following notation, N_g is the set of events occur, λ_g is the intensity and f is the so-called *mark's distribution*. The * symbolises that the the intensity and mark distribution depend on previous events. Namely, the intensity at time t_i depends on the events, $\{(t_1, \kappa_1), \ldots, (t_{i-1}, \kappa_{i-1})\}$. Also note the distinction that $\lambda(t_i)$ does not depend on the mark κ_i, but the mark is allowed to depend on time t_i. The subscript g is a common notation used to imply this is the intensity of the ground process, i.e. we are only considering the events of interest. Further details of likelihoods of MPP can be found in [DVJ03, Section 7.3].

One reason that we believe MPPs are a good choice for this particular problem is that one can view them as natural generalisation of CTMCs. This is apparent from the likelihood since, letting $\lambda = q_i$ and $f = q_{ij}/q_i$ we recover the likelihood of a CTMC, see (5.1.7).

Chapter 6

In the Markov Setting

The current regulation modelling assumption is that a company's rating transition is Markov. That is, only the company's current rating determines its transition probabilities. Although this assumption has been called into question (see [LS02] and Section 7.1 for example), this does make calculations more tractable. We discuss removing the Markov assumption in Chapter 7.

Recalling the problem. We take the view of a financial agent who wishes to estimate probabilities of default or assess risk in their portfolio due to credit transitions but does not have access to (the expensive) individual credit rating transitions. The agent only has the annual TPM (missing data case), say $P(1)$, and uses a continuous time Markov chain (CTMC), say $(\hat{P}(t))_{t\geq 0}$, with a finite state space to model the changes in rating over time. Under standard conditions the evolution of the CTMC can be written as $\hat{P}(t) = e^{Qt}$ where Q is the generator matrix. The problem is then to estimate Q given $P(1)$.

This estimation is non-trivial due to the so-called embeddability problem (as mentioned in Chapter 1). It is discussed in great detail by [IRW01] and, for further discussion and results, we point the reader to [Lin11].

Several approaches exist to tackle this estimation problem [KS01, IRW01, TÖ04, BS05, Ina06, BS09], either using deterministic algorithms (e.g. diagonal or weighted adjustment, Quasi-optimization of the generator) or statistical ones (Expectation-Maximisation (EM), Markov chain Monte-Carlo (MCMC) ones), see Section 6.2. We focus on the Expectation-Maximisation algorithm of [BS05] for CTMCs and allow for an absorbing states.

Remark 6.0.1. *If a matrix P is embeddable*[*]*, the algorithms below are pointless and one can easily tackle the problem through eigenvalue decomposition etc. Or in the case where the exact timing of rating transitions are known (full data case) one can use the standard maximum likelihood estimator as in* [JLT97]. *In this chapter the only data given is a set of yearly TPMs which in general are not embeddable and the methods just mentioned do not yield useful results.*

Preliminaries and standing convention. Throughout this part we consider companies defined on a finite state space $\{1, \dots, h\}$, where each state corresponds to a rating. We denote AAA as rating 1 and D (default) as rating h. We adopt the standard notation that P is an h-*by*-h stochastic matrix, which will be the observed TPM (at, say, time $t = 1$) and Q is an h-*by*-h generator matrix. We further denote by $P_{ij} := (P)_{ij}$, by $q_{ij} := (Q)_{ij}$ and the intensity of state i by $q_i = \sum_{j \neq i} q_{ij}$ where $i, j \in \{1, \dots, h\}$. A standard assumption used in credit risk modelling is that default is an absorbing state, hence $P_{hh} = 1$ (which is the case when we set $q_h = 0$). We work with infinitesimal generators of the following class.

Our Contribution Firstly we provide sufficient conditions to extend the convergence result of [BS05] to individual parameters rather than just convergence of likelihoods. The conditions presented are trivially satisfied in the context of the TPM problem. Secondly, we derive two closed form expressions for the entries of the Hessian of the likelihood function used in the EM algorithm. This eliminated several instability issues appearing in other numerical implementations found in the literature and allows for computational speedups (comparatively). Moreover, the result provides a way to estimate the error of the estimation via the Fisher information matrix (Wald

[*] In this setting a stochastic matrix P is embeddable if there exists a generator Q such that $P = e^{Q}$.

intervals) and assess the nature of the stationary point the algorithm has converged to. We further provide expressions allowing one to transfer confidence intervals at the level of the generator matrix to the level of rating transitions and probabilities of default, where they can be easily interpreted. Finally we give a short overview of known methods and implement them with some modifications as to improve their performance. See Sections 6.2 & 6.3 for precise meanings: we apply the algorithms to certain credit risk problems and carry out a simulation study to check the impact in the computation of *risk charges*, namely IRC (Incremental Risk Charge) with VaR (Value at Risk), IDR (Incremental Default Risk) with VaR and IRC with ES (Expected Shortfall). We distinguish portfolio types (mixed, investment or speculative); the impact of different types of generators (stable vs unstable); dependence on the sample size and general convergence. We compare probabilities of default as maps of time across different algorithms and find interesting results.

This chapter is organised as follows. In Section 6.1 we present the EM algorithm and we state our main theoretical findings, in particular in Section 6.1.2 and 6.1.3 we establish our closed form expressions for the Wald confidence intervals for the generator and the underlying TPM. In Section 6.2 we briefly present other known algorithms and in Section 6.3 we present the benchmarking results. Finally we calculate the Wald confidence intervals on empirical data in Section 6.4.

Remark 6.0.2 (Software availability). *The findings and algorithms of this work are now part of an improved version of the CRAN R-package* ctmcd: Estimating the Parameters of a Continuous-Time Markov Chain from Discrete-Time Data *(see* [Pfe17]*) — https://CRAN.R-project.org/package=ctmcd*

6.1 The EM Algorithm

There exists extensive literature on the majority of the algorithms we present in this section, therefore we only provide brief discussions and include references for additional information. Further, we will use the theory of Markov chains extensively.

If a matrix Q satisfies the conditions in Definition 5.1.2, then for all $t \geq 0$ the matrix $P(t) := e^{Qt}$ is a stochastic matrix (Theorem 5.1.3), where e^{A} is the matrix exponential of matrix A. The goal of the algorithms presented is to calculate a generator matrix Q such that e^{Qt} is the "best fit" to the observed TPM, where t denotes the length of time between the rating updates (typically one year).

As in Section 5.1, let $(X(t))_{t\geq 0}$ denote a CTMC over the finite state space $\{1, \ldots, h\}$ with a generator Q of the above class. Associated to $X(t)$ is, for i, j in the state space, $K_{ij}(t)$ the number of jumps from i to j in the interval $[0, t]$ and by $S_i(t)$ the holding time of state i in the interval $[0, t]$.

6.1.1 The Algorithm

Many methods have been developed in statistics in order to obtain the maximum likelihood estimate, but many methods break in the presence of *missing data*. Mathematically, we are interested in some set $\mathcal{X}$, but we are only able to observe $\mathcal{Y}$, with the assumption there is a many-to-one mapping from $\mathcal{X}$ to $\mathcal{Y}$. That is, $\mathcal{X}$ is a much richer set than $\mathcal{Y}$.

When dealing with such a case, the Expectation Maximisation (EM) algorithm often offers a robust solution to the problem. [MK07] provide a complete overview of the algorithm. The basis of the algorithm is, we observe data y which is a realisation (element) of $\mathcal{Y}$. We know y has density function g (sometimes referred to as a sampling density) depending on parameters Ψ in some space Λ, but we want the density (likelihood) of $\mathcal{X}(y)$. Hence, postulate some family of densities f, dependent on Ψ, where f corresponds to the density of the complete data set $\mathcal{X}(y)$ (the set of points $x \in \mathcal{X}$ which are in the pre-image of $y \in \mathcal{Y}$). The relation between f and g is,

$$g(y; \Psi) = \int_{\mathcal{X}(y)} f(x; \Psi) \mathrm{d}x \,.$$

The idea is, the EM algorithm maximises g w.r.t. Ψ, but we force it to do so by using the density f. Further, define,

$$R(\Psi';\Psi) := \mathbb{E}_\Psi\big[\ln\big(f(x;\Psi')\big)\big|y\big] \qquad \text{for } \Psi',\Psi\in\Lambda, \tag{6.1.1}$$

where $\mathbb{E}_\Psi[\cdot|y]$ is the conditional expectation, conditional on y under parameters Ψ. We assume R to exist for all pairs (Ψ',Ψ), in particular we assume $f(x;\Psi) > 0$ almost everywhere in $\mathcal{X}$ for all Ψ (otherwise the logarithm is infinite). Let us clarify, f is calculated using Ψ', but the expectation is calculated using Ψ. The EM algorithm is then the following iterative procedure.

1. Choose an initial $\Psi^{(1)}$ and take $p = 1$.

2. E-step: Compute $R(\Psi;\Psi^{(p)})$.

3. M-step: Choose $\Psi^{(p+1)}$ to be the value of $\Psi\in\Lambda$ that maximises $R(\Psi;\Psi^{(p)})$.

4. Check if the predefined convergence criteria is met, if not, take $p = p+1$ and return to (ii).

The particular problem of generator estimation For our problem the observed process is a discrete time Markov chain (DTMC), the unobserved process to estimate is a continuous time Markov chain (CTMC). Therefore, the observed data is the discrete transitions (annual TPM) and the parameters we wish to estimate are the entries in the generator. Mathematically we consider the case where the CTMC is observed at times $t_0 < t_1 < \cdots < t_M$, denote by $\Delta t_u := t_u - t_{u-1}$ for $u\in\{1,\dots,M\}$ and the transition matrix over that interval by $N(u)$. The likelihood of the discretely observed CTMC is given by (the g in the notation above),

$$L(Q|N) = \prod_{u=1}^{M}\prod_{s=1}^{h}\prod_{r=1}^{h}\exp(Q\Delta t_u)_{ij}^{N_{sr}(u)}. \tag{6.1.2}$$

Even though this is not the likelihood of a CTMC, it is the likelihood based on what we can observe so in effect the EM algorithm looks to find Q to maximise (6.1.2). Therefore the Wald confidence intervals are also based on this likelihood.

Recall from Section 5.1, the likelihood of a fully observed CTMC is (f in the notation above)

$$L_t(Q) = \exp\Big(\sum_{i=1}^{h}\Big[\sum_{j\neq i}\log(q_{ij})K_{ij}(t) - S_i(t)\sum_{j\neq i}q_{ij}\Big]\Big).$$

Hence given two generators Q', Q, the function R in (6.1.1) is,

$$R(Q';Q) = \sum_{i=1}^{h}\left[\sum_{j\neq i}\log(q'_{ij})\mathbb{E}_Q[K_{ij}(t)|y] - \mathbb{E}_Q[S_i(t)|y]\sum_{j\neq i}q'_{ij}\right], \tag{6.1.3}$$

where y denotes the discrete time observations. This gives important intuition about the EM algorithm, recall that the issue with only having the discrete time data is that one does not know K and S (missing data), hence likelihood maximisation is more difficult. The purpose of the EM is then to replace the data we do not have access to by the corresponding expected value. Maximising for q'_{ij} in $R(Q';Q)$ yields

$$q'_{ij} = \frac{\mathbb{E}_Q[K_{ij}(t)|y]}{\mathbb{E}_Q[S_i(t)|y]}. \tag{6.1.4}$$

The difficult step is the calculation of $\mathbb{E}_Q[K_{ij}(t)|y]$ and $\mathbb{E}_Q[S_i(t)|y]$. We follow an approach similar to [BS05] (see also [BMNS02]) but express the result in a framework more suited to generator estimation from TPMs, rather than the estimation from individual movements. Furthermore, the result derived in [BMNS02] is for irreducible Markov chains making it not applicable to our case (CTMC with absorbing states), accounted for in Proposition 6.1.2.

Consider the following functions (see [BMNS02]), for $1 \leq i, j \leq h$

$$V^*_{ij}(c, Z; t) = \mathbb{E}_Q \left[\exp\Big(- \sum_{\mu=1}^{h} c_\mu S_\mu(t) \Big) \prod_{\mu,\nu=1}^{h} Z_{\mu\nu}^{K_{\mu\nu}(t)} \mathbb{1}_{\{X(t)=j\}} \middle| X(0) = i \right] , \qquad (6.1.5)$$

where we denote by $c = (c_1, \cdots, c_h) \in \mathbb{R}^h$ and $Z \in \mathbb{R}^{h \times h}$ with $Z_{ii} = 1$ for $i \in \{1, \cdots, h\}$. Observe that V^*_{ij} is the Laplace-Stieltjes transform of the holding times S and the probability generating function of the jumps K, with initial and final states $X(0) = i$ and $X(t) = j$ respectively. Denoting by $V^*(c, Z; t)$ the h-*by*-h matrix of elements $V^*_{ij}(c, Z; t)$. This allows us to give the main theorem (similar version) in [BMNS02].

Theorem 6.1.1. *For $t \geq 0$, the matrix $V^*(c, Z; t)$ is given by,*

$$V^*(c, Z; t) = \exp\big([Q \bullet Z - \Delta(c)]t\big) ,$$

where $\bullet$ is the Schur (Hadamard) product[†] *of matrices, Q is the generator matrix from the CTMC and $\Delta(c)$ is the diagonal matrix with entries c_i at position ii for $i = 1, \ldots, h$.*

A closed form expression for the expectation terms in (6.1.4) follows from a result in [VL78] (sketched also in [HJ11]).

Proposition 6.1.2. *Let e_i be the column vector of length h which is one at entry i and zero elsewhere, further let us define the $2h$-by-$2h$ matrices $C_\gamma^{(\alpha\beta)}$ and $C_\phi^{(\alpha)}$ as,*

$$C_\gamma^{(\alpha\beta)} = \begin{bmatrix} Q & q_{\alpha\beta} e_\alpha e_\beta^\intercal \\ 0 & Q \end{bmatrix} \qquad \textit{and} \qquad C_\phi^{(\alpha)} = \begin{bmatrix} Q & e_\alpha e_\alpha^\intercal \\ 0 & Q \end{bmatrix} \quad \alpha, \beta \in \{1, \cdots, h\}. \qquad (6.1.6)$$

Consider a CTMC X that we observe at n time points $0 \leq t_1 < t_2 < \cdots < t_n$ and denote by y_s the state of the Markov chain at time t_s, i.e. $y_s := X(t_s)$. Then, the expected jumps and holding times over the observations are,

$$\mathbb{E}_Q[K_{ij}(t)|y] = \sum_{s=1}^{n-1} \frac{\Big(\exp\big(C_\gamma^{(ij)}(t_{s+1} - t_s)\big)\Big)_{y_s, h+y_{s+1}}}{\big(\exp\big(Q(t_{s+1} - t_s)\big)\big)_{y_s, y_{s+1}}} ,$$

$$\mathbb{E}_Q[S_i(t)|y] = \sum_{s=1}^{n-1} \frac{\Big(\exp\big(C_\phi^{(i)}(t_{s+1} - t_s)\big)\Big)_{y_s, h+y_{s+1}}}{\big(\exp\big(Q(t_{s+1} - t_s)\big)\big)_{y_s, y_{s+1}}} .$$

Proof. We firstly note that a similar technique to the one below is detailed in [HJ11], however, for completeness we detail the steps here. Observe that V^* in (6.1.5) satisfies the differential equation in the statement of Theorem 6.1.1 (see [BS05]).

$$\frac{\mathrm{d}}{\mathrm{d}t} V^*_{\mu\nu}(c, Z; t) = -V^*_{\mu\nu}(c, Z; t) c_\nu + \sum_{r=1}^{h} V^*_{\mu r}(c, Z; t) q_{r\nu} Z_{r\nu} \qquad \mu,\ \nu \in \{1, \ldots, h\}. \qquad (6.1.7)$$

To simplify the presentation, define for any two states μ, $\nu \in \{1, \ldots, h\}$, satisfying the *positive probability condition* $\mathbb{P}_Q[X(t) = \nu | X(0) = \mu] > 0$, for $t > 0$,

$$\bar{\xi}^{ij}_{\mu\nu}(t) := \mathbb{E}_Q[K_{ij}(t) | X(t) = \nu, X(0) = \mu] \quad \text{and} \quad \bar{\zeta}^{i}_{\mu\nu}(t) := \mathbb{E}_Q[S_i(t) | X(t) = \nu, X(0) = \mu] .$$

Note that the *positive probability condition* allows for the Markov process to be reducible. It is a trivial calculation to show that,

$$\mathbb{E}_Q[K_{ij}(t)|y] = \sum_{s=1}^{n-1} \bar{\xi}^{ij}_{y_s y_{s+1}}(t_{s+1} - t_s) \quad \text{and} \quad \mathbb{E}_Q[S_i(t)|y] = \sum_{s=1}^{n-1} \bar{\zeta}^{i}_{y_s y_{s+1}}(t_{s+1} - t_s) . \qquad (6.1.8)$$

[†] The Shur product of two $h \times h$ matrices A and B is the $h \times h$ matrix with elements $A_{ij} B_{ij}$.

Before continuing we study the related quantities,

$$\xi^{ij}_{\mu\nu}(t) = \mathbb{E}_Q[K_{ij}(t)\mathbb{1}_{\{X(t)=\nu\}}|X(0)=\mu] \quad \text{and} \quad \zeta^{i}_{\mu\nu}(t) = \mathbb{E}_Q[S_i(t)\mathbb{1}_{\{X(t)=\nu\}}|X(0)=\mu]\,.$$

Again we only consider indices μ, ν s.t. $\mathbb{P}_Q[X(t)=\nu|X(0)=\mu]>0$, for $t>0$. From standard conditional probability, the relationship between these quantities is,

$$\bar{\xi}^{ij}_{\mu\nu}(t) = \frac{\xi^{ij}_{\mu\nu}(t)}{\mathbb{P}_Q[X(t)=\nu|X(0)=\mu]} \quad \text{and} \quad \bar{\zeta}^{i}_{\mu\nu}(t) = \frac{\zeta^{i}_{\mu\nu}(t)}{\mathbb{P}_Q[X(t)=\nu|X(0)=\mu]}\,. \tag{6.1.9}$$

By defining $\xi^{ij}(t)$ as the h-*by*-h matrix with μ, ν entry $\xi^{ij}_{\mu\nu}(t)$, where we define $\xi^{ij}_{\mu\nu}(t)=0$ for μ, ν such that $\mathbb{P}_Q[X(t)=\nu|X(0)=\mu]=0$. It is shown in [BS05] that ξ^{ij} satisfies the following differential equation, with the obvious condition $\xi^{ij}(0)=0$,

$$\frac{\mathrm{d}}{\mathrm{d}t}\xi^{ij}(t) = q_{ij}\exp(Qt)e_i e_j^\intercal + \xi^{ij}(t)Q$$

By considering a system of linear inhomogeneous differential equations its solution is given by (see [Tes12, Chapter 3.3]),

$$\xi^{ij}(t) = \int_0^t \exp(Qs)q_{ij}e_i e_j^\intercal \exp((t-s)Q)\mathrm{d}s\,. \tag{6.1.10}$$

Through a similar argument we obtain ζ^i as,

$$\zeta^{i}(t) = \int_0^t \exp(Qs)e_i e_i^\intercal \exp((t-s)Q)\mathrm{d}s\,. \tag{6.1.11}$$

As it turn out we can obtain explicit solution to this integral using the method in [VL78] as follows (this technique is described in [HJ11]), consider the upper triangular matrix,

$$C = \begin{bmatrix} A_1 & B_1 \\ 0 & A_2 \end{bmatrix},$$

where A_1, A_2 and B_1 are h-*by*-h matrices. Noting that any integer power of an upper triangular matrix is also upper triangular. Then the exponential of C is an upper triangular matrix, hence,

$$\exp(Ct) = \begin{bmatrix} F_1(t) & G_1(t) \\ 0 & F_2(t) \end{bmatrix}.$$

Recalling for matrix exponentials, $\frac{\mathrm{d}}{\mathrm{d}t}e^{Ct} = Ce^{Ct} = e^{Ct}C$. Hence we can define,

$$\frac{\mathrm{d}}{\mathrm{d}t}e^{Ct} = \begin{bmatrix} F_1(t) & G_1(t) \\ 0 & F2(t) \end{bmatrix}\begin{bmatrix} A_1 & B_1 \\ 0 & A_2 \end{bmatrix} = \begin{bmatrix} F_1(t)A_1 & F_1(t)B_1 + G_1(t)A_2 \\ 0 & F_2(t)A_2 \end{bmatrix}.$$

Alternatively, we can write, $\dot{F}_j(t) = F_j(t)A_j \implies F_j(t) = \exp(A_j t)$, for $j=1,2$ and $\dot{G}_1(t) = F_1(t)B_1 + G_1(t)A_2$. Again using the same result for inhomogeneous differential equations we may write the solution to the differential equation $\dot{G}_1(t) = F_1(t)B_1 + G_1(t)A_2$ as,

$$G_1(t) = \int_0^t F_1(s)B_1 e^{-sA_2}\mathrm{d}s\, e^{tA_2} = \int_0^t e^{sA_1}B_1 e^{(t-s)A_2}\mathrm{d}s\,.$$

Setting $A_1 = A_2 = Q$ and $B_1 = q_{ij}e_i e_j^\intercal$, yields

$$G_1(t) = \int_0^t e^{sQ}q_{ij}e_i e_j^\intercal e^{(t-s)Q}\mathrm{d}s = \text{(6.1.10)}.$$

Hence, taking C_γ as defined in (6.1.6), one obtains

$$\exp(C_\gamma^{ij}t) = \left[\begin{array}{cc} e^{Qt} & \xi^{ij}(t) \\ 0 & e^{Qt} \end{array}\right] \quad \Longrightarrow \quad \xi^{ij}(t) = \big(\exp(C_\gamma^{ij}t)\big)_{1:h,h+1:2h}.$$

Similarly, setting $A_1 = A_2 = Q$ and $B_1 = e_i e_i^\intercal$, produces (6.1.11). By then using the fact that $\mathbb{P}_Q[X(t) = \nu | X(0) = \mu] = (\exp(Qt))_{\mu\nu}$ along with (6.1.9) and (6.1.8) we obtain the required results. □

Thus we obtain closed form expressions for the two key quantities appearing in (6.1.4). This approach differs from [BS05] where they describe numerical schemes to solve the differential equations, namely Runge-Kutta and uniformization. These techniques can yield good results at this level, but our closed form expression will pay dividends when it comes to error estimation.

This yields the relation we desire, however, in our example we have an observed TPM (or sequence of TPMs), P, in the case of equal observation windows, t in the interval $[0, T]$ (although it is trivial to generalise) the expectation can be expressed as,

$$\begin{aligned} \mathbb{E}_Q[K_{ij}(T)|P] &= \sum_{u=1}^{N}\sum_{s=1}^{h}\sum_{r=1}^{h} P_{sr}^u(t) \frac{\left(\exp\left(C_\gamma^{(ij)}t\right)\right)_{s,h+r}}{(\exp(Qt))_{s,r}}, \\ \mathbb{E}_Q[S_i(T)|P] &= \sum_{u=1}^{N}\sum_{s=1}^{h}\sum_{r=1}^{h} P_{sr}^u(t) \frac{\left(\exp\left(C_\phi^{(i)}t\right)\right)_{s,h+r}}{(\exp(Qt))_{s,r}}, \end{aligned} \qquad (6.1.12)$$

where $n = T/t$ (the number of observations) and P^u is the TPM of the u-th observation.

Remark 6.1.3 (The reducible case). *Previously, we only had observed transitions, hence they must have a non-zero probability of occurring. Here we can sum s and r over the full range because $P_{sr}(t)$ acts as an indicator of possible transitions, that is, if $P_{sr}(t) = 0$ we set the s, r component as 0. Clearly, if $P_{sr}(t) > 0$, but $(\exp(Qt))_{sr} = 0$, Q is misspecified.*

Roughly speaking, the above formula is taking each row in the TPM to contain equal amounts of information (observations). When one knows the number of transitions between the states N, then one replaces $P_{sr}^u(t)$ by $N_{sr}(u)$, where $N_{sr}(u)$ is the number of observed transitions in observation u.

Likelihood Convergence of the EM algorithm In the case of this problem [BS05] provide a proof that the likelihood function converges with one small caveat in order to keep the parameter space compact. Namely, they use the following constrained parameter space, $\mathcal{Q}_\epsilon$, which can be achieved by setting, $\mathcal{Q}_\epsilon = \{Q \in \mathcal{Q} | \det[\exp(Q)] \geq \epsilon\}$ ($\mathcal{Q}$ is the parameter space from Definition 5.1.2) for some $\epsilon > 0$. Theorem 4 in [BS05] states that the algorithm will converge to a stationary point of the likelihood or hit the boundary of the parameter space they have induced. It is accepted this is a crude approach to solving the problem and further analysis is needed when $\det[\exp(Q)] = \epsilon$. An alternative approach would be to use a penalised likelihood as discussed in [MK07, p.214].

Parameter convergence criteria The above convergence is sufficient for one to conclude convergence of the likelihood. However, it is not sufficient for convergence of the parameters as one cannot state that the series of iterates $Q^{(k)}$ converge ($\|Q^{(k+1)} - Q^{(k)}\| \to 0$ as $k \to \infty$). From a theoretical standpoint this may not be as important as convergence of the likelihood itself, nonetheless, it is of key importance for applications. For instance, without convergence of the parameters the risk charge different financial agents obtain from the same data may vary wildly, even under very strict convergence conditions. Before proving convergence we require two important points.

Remark 6.1.4. *With* (6.1.12) *in mind we assume that for any $s \neq r$ such that $P_{sr}^u(t) = 0$ for all u, we take the starting point $q_{sr}^{(0)} := (Q^{(0)})_{sr} = 0$. As discussed in* [BS05], *any point set to zero will* *stay at zero for all iterations. Note, we are not changing the problem since these terms will converge to zero under the EM algorithm.*

Assumption 6.1.5 (Element constraint)**.** *Similar to* [BS05]*, we will use a manual space constraint to obtain the convergence. Take $1 > \epsilon > 0$, such that $\forall\ i \neq j$, $q_{ij} < 1/\epsilon$. Moreover, we assume adjacent mixing, namely, for $i \in \{2, \dots, h-1\}$, $q_{i,i\pm1} > \epsilon$ and $q_{1,2} > \epsilon$.*

We denote the space of generator matrices which satisfy this condition as Λ_ϵ.

The above assumption ensures non-zero entries in the tri-diagonal band and also only finite entries as one can take ϵ as small as we wish. In the case of TPMs associated to credit ratings, such an assumption is trivially satisfied as one generally has diagonally dominant matrices and companies can always be upgraded or downgraded by one, thus $P^u_{i,i\pm1}$ are typically non-zero. Diagonal dominance is sufficient for the generator to be identifiable and therefore entries do not blow up, we discuss the notion of identifiability in Section 6.1.2.

Proving the parameters converge is more challenging than the likelihoods, however, [Wu83] provide a sufficient condition for this to occur, namely a sufficient condition for $\|Q^{(k+1)} - Q^{(k)}\| \to 0$ as $k \to \infty$ is, there exists a forcing function[‡] F such that,

$$R(Q^{(k+1)}; Q^{(k)}) - R(Q^{(k)}; Q^{(k)}) \geq F(||Q^{(k+1)} - Q^{(k)}||)\,.$$

An example of a forcing function is $\sigma(t) = \lambda t^2$ where $\lambda > 0$. We require the following bounds on the expected values to show convergence.

Lemma 6.1.6. *Let n and P^u be as defined in* (6.1.12) *and assume for $i \neq j$ there exists a $u \in \{1, \dots, n\}$ such that $P^u_{ij} > 0$ (we observe a movement from i to j in observation window u). Then we obtain the following bounds on the expected number of jumps:*

$$P^u_{ij}\frac{\epsilon q_{ij}}{h} \leq \mathbb{E}_Q[K_{ij}(T)|P] \leq h^2 n \frac{ht}{\epsilon \min\left\{\epsilon^h t^h \exp(-th^2/\epsilon)\ ,\ \exp(ht/\epsilon)\right\}}\,. \tag{6.1.13}$$

Moreover, assuming there exists a $u \in \{1, \dots, N\}$ such that $P^u_{ii} > 0$, we obtain the following bound on the expected holding time,

$$\mathbb{E}_Q[S_i(T)|P] \geq P^u_{ii} t \exp\Big(-\frac{ht}{\epsilon}\Big)\,. \tag{6.1.14}$$

To maintain the flow of the text we state immediately our main convergence result, and defer the proof of the Lemma to Section 6.5.1.

Theorem 6.1.7. *Under Assumption 6.1.5, then, there exists a $\lambda > 0$ such that for all EM iterations $k \in \mathbb{N}$,*

$$R\big(Q^{(k+1)}; Q^{(k)}\big) - R\big(Q^{(k)}; Q^{(k)}\big) \geq \lambda \|Q^{(k+1)} - Q^{(k)}\|^2\,,$$

where $||\cdot||$ is the Euclidean norm.

Proof. Writing out the difference in the R terms we obtain,

$$\sum_{i=1}^{h}\sum_{j\neq i}\Big[\mathbb{E}_{Q^{(k)}}[K_{ij}(t)|P]\big(\log(q^{(k+1)}_{ij}) - \log(q^{(k)}_{ij})\big) - \mathbb{E}_{Q^{(k)}}[S_i(T)|P]\big(q^{(k+1)}_{ij} - q^{(k)}_{ij}\big)\Big]\,.$$

Due to the form of the Euclidean norm squared and the function R, it is sufficient to show the inequality holds for all $i \neq j$. Namely, it is sufficient to show the existence of a $\lambda > 0$ such that,

$$\mathbb{E}_{Q^{(k)}}[K_{ij}(T)|P]\big(\log(q^{(k+1)}_{ij}) - \log(q^{(k)}_{ij})\big) - \mathbb{E}_{Q^{(k)}}[S_i(T)|P]\big(q^{(k+1)}_{ij} - q^{(k)}_{ij}\big) \geq \lambda\big(q^{(k+1)}_{ij} - q^{(k)}_{ij}\big)^2, \tag{6.1.15}$$

for all $i \neq j$. We tackle the log terms first. It is well known that we can express any C^∞-function using Taylor expansion to a finite number of terms with some error (remainder) term. Moreover,

[‡] A forcing function is defined as any function $F : [0, \infty) \to [0, \infty)$ such that for any sequence t_k defined in $[0, \infty)$, $\lim_{k\to\infty} F(t_k) = 0$ implies $\lim_{k\to\infty} t_k = 0$.

the error term has a known form and hence, using an exact Taylor expansion to second order, there exists a $Z \in [\min(q_{ij}^{(k)}, q_{ij}^{(k+1)}), \max(q_{ij}^{(k)}, q_{ij}^{(k+1)})]$ such that,

$$\log(q_{ij}^{(k+1)}) - \log(q_{ij}^{(k)}) = \frac{-1}{q_{ij}^{(k+1)}}(q_{ij}^{(k)} - q_{ij}^{(k+1)}) + \frac{1}{2Z^2}(q_{ij}^{(k)} - q_{ij}^{(k+1)})^2 \,,$$

where we have expanded $q_{ij}^{(k)}$ around $q_{ij}^{(k+1)}$. Substituting (6.1.4) into the LHS of (6.1.15), the condition simplifies to,

$$\frac{\mathbb{E}_{Q^{(k)}}[K_{ij}(T)|P]}{2Z^2}(q_{ij}^{(k)} - q_{ij}^{(k+1)})^2 \geq \lambda(q_{ij}^{(k+1)} - q_{ij}^{(k)})^2 \,.$$

In order to show this bound we need to get a handle on Z. Clearly, there are two options between iterations, either $q_{ij}^{(k)} > q_{ij}^{(k+1)}$ or $q_{ij}^{(k)} \leq q_{ij}^{(k+1)}$. In the latter case we obtain,

$$\frac{\mathbb{E}_{Q^{(k)}}[K_{ij}(T)|P]}{2Z^2}(q_{ij}^{(k)} - q_{ij}^{(k+1)})^2 \geq \frac{\mathbb{E}_{Q^{(k)}}[S_i(T)|P]^2}{2\mathbb{E}_{Q^{(k)}}[K_{ij}(T)|P]}(q_{ij}^{(k)} - q_{ij}^{(k+1)})^2 \,.$$

Since we can element wise bound $Q^{(k)}$, using Lemma 6.1.6 and Assumption 6.1.5 we can bound the term $\mathbb{E}_{Q^{(k)}}[K_{ij}(T)|P]$ from above and $\mathbb{E}_{Q^{(k)}}[S_i(T)|P]$ from below by constants (depending on ϵ). Hence, we can choose a λ independent of k such that the condition is satisfied.

The second case $q_{ij}^{(k)} > q_{ij}^{(k+1)}$, follows a similar argument. Again, we can set Z as the larger of the two values, thus we obtain the following inequality,

$$\frac{\mathbb{E}_{Q^{(k)}}[K_{ij}(T)|P]}{2Z^2}(q_{ij}^{(k)} - q_{ij}^{(k+1)})^2 \geq \frac{\mathbb{E}_{Q^{(k)}}[K_{ij}(T)|P]}{2\big(q_{ij}^{(k)}\big)^2}(q_{ij}^{(k)} - q_{ij}^{(k+1)})^2 \,.$$

Using Lemma 6.1.6, we can reduce this inequality to,

$$\frac{\mathbb{E}_{Q^{(k)}}[K_{ij}(T)|P]}{2Z^2}(q_{ij}^{(k)} - q_{ij}^{(k+1)})^2 \geq \frac{P_{ij}^u \epsilon}{2hq_{ij}^{(k)}}(q_{ij}^{(k)} - q_{ij}^{(k+1)})^2 \,.$$

Since $P_{ij}^u > 0$ and we can bound each q_{ij} from above, again we choose a λ independent of k. □

Starting value for the EM algorithm The final point to discuss, is the choice of the initial matrix Q. It is useful from a computational point of view to start in a good place. Here we choose Q based on a generalisation of the QOG algorithm (described in Section 6.2.1) that allows for complex inputs. For each entry q_{ij} we define the input as,

$$q_{ij} \to \text{sign}\big(\text{Re}(q_{ij})\big) \times |q_{ij}| \,,$$

where $|q_{ij}|$ is the magnitude of q_{ij} and Re(q_{ij}), is the real component of q_{ij}. With the newly defined Q we apply the QOG algorithm. We take any zero entries not in the final row to be a small number (10^{-5}, say) unless there are zero observed transitions. This defines our initial choice of Q. We define the EM algorithm steps as,

1. Take an initial intensity matrix Q and positive value ϵ.
2. While the convergence criteria is not met and all entries of Q are within the boundaries
   (1) E-step: calculate $\mathbb{E}_Q[K_{ij}(T)|P]$ and $\mathbb{E}_Q[S_i(T)|P]$.
   (2) M-step: set $q'_{ij} = \mathbb{E}_Q[K_{ij}(T)|P]/\mathbb{E}_Q[S_i(T)|P]$, for all $i \neq j$ and set q_{ii} appropriately.
   (3) Set $Q = Q'$ (where Q' is the matrix of q's) and return to E-step.
3. End while and return Q.

This leads to the following theorem for convergence of the EM.

Theorem 6.1.8 (Convergence of the EM). *Assume that our initial point is in the parameter space Λ_ϵ: is a true generator and satisfies Assumption 6.1.5. Then either the sequence of points $\{Q^{(k)}\}_k$ converges to a single point in Λ_ϵ which is also a stationary point of the likelihood, or the entries go to the boundary (blow up or some tri-diagonal elements in an non-absorbing row go to zero).*

A proof of Theorem 6.1.8 follows directly from Theorem 4 in [BS05] and our Theorem 6.1.7.

Remark 6.1.9 (The unique maximiser of the Likelihood). *The natural question one may ask is does this stronger form of convergence imply convergence to the global maximum? The problem of existence and uniqueness of maximum likelihoods in this setting is a very challenging problem with a long history.* [BS05] *give a wonderful overview on the subject, Theorem 1 in* [BS05] *also provides results on existence and uniqueness of the maximum. Unfortunately, one cannot say more than this, if one can derive conditions under which a unique maximum existed (for non-embeddable TPMs) then the above convergence result is sufficient to conclude the EM will converge to the MLE.*

Our Theorem 6.1.7 is handy in this context as it shows that once the EM lands "near" the global maximum the iteration will converge to it.

6.1.2 Variance Estimation

In this section we derive an expression for the Hessian of the likelihood. We use a result in [Oak99] and follow [BS09], however, unlike [BS09], we provide a closed form expression for the Hessian. This result eliminates the stability problems observed in the numerical simulation case when the entries in Q are small. The Hessian provides a way to estimate the Wald confidence intervals of the maximum likelihood estimates and further allows us to assess the nature of the converged stationary point (this is further discussed in Section 6.4.1).

We point the reader to [BS05, Theorem 1] for results on the existence and uniqueness of maximum likelihood estimators with respect to this problem. Further, for discussions on consistency and asymptotic normality related to this problem one should consult [KW13], [KW14]. [KW13], provide sufficient conditions for consistency, the key assumption relies on so-called model *identifiability*[§]. [KW13] prove *identifiability* under conditions which are too restrictive for our purpose; [BS05, DY07] discusses the problem of *identifiability* in detail. From [Cut73], [BS05] for the model to be identifiable it is sufficient (though very crude) to have $\min_i(\exp(Qt))_{ii} > 1/2$, [Cul66] gives a requirement for general matrices based on the eigenvalues which one can always aposteriori verify after a Q is deduced. The crucial assumption in [KW14] to obtain asymptotic normality, is that the Hessian must be invertible at the true value, we can of course verify invertibility a posteriori.

Let us now state the following definition.

Definition 6.1.10 (Allowed pairs). *We say that the pair α, β is allowed if $\alpha \neq \beta$ (not in the diagonal) and $q_{\alpha\beta}$ is not converging to zero under the EM algorithm.*

For practical applications, one can imagine the set of allowed values, as the set of α, β such that $q_{\alpha\beta} > \epsilon$, where ϵ is some cut-off point (10^{-8}, say). The reason we must exclude small parameters is, this analysis only holds in the large data limit, since we do not have an infinite amount of data we cannot for certain rule out some jump, however, if $q_{\alpha\beta}$ is converging to zero, it implies that this parameter is either zero or extremely close to zero and therefore we can bound it above by a small number. Moreover, from a mathematical point of view a parameter which does tend to zero (or even becomes zero) lies on the boundary, where the notion of differentiability is not clear. Therefore, we can think of the "allowed pairs" as the variables when solving the problem in a restricted parameter space.

As it turns there are two approaches to obtain the Hessian for the EM in this setting, which we now detail both.

Hessian via Oakes formula

The first approach is to use the result from [Oak99] for calculating the Hessian of the likelihood, in the likelihood.

[§] In our setting a model is identifiable if there does not exist a generator $Q' \neq Q$ such that $\exp(Qt) = \exp(Q't)$.

Lemma 6.1.11. *The second derivative of the likelihood with parameter Ψ and observed information y is related to the EM function R by*

$$\frac{\partial^2 L(\Psi; y)}{\partial \Psi^2} = \left[\frac{\partial^2 R(\Psi'; \Psi)}{\partial \Psi'^2} + \frac{\partial^2 R(\Psi'; \Psi)}{\partial \Psi' \partial \Psi}\right]_{\Psi' = \Psi} .$$

Injecting (6.1.3) in the above we obtain,

$$\frac{\partial^2 R(Q'; Q)}{\partial q'_{\alpha\beta} \partial q'_{\mu\nu}} = \frac{-1}{q'^2_{\mu\nu}} \mathbb{E}_Q[K_{\mu\nu}(t)|y] \delta_{\alpha\mu} \delta_{\beta\nu} , \tag{6.1.16}$$

$$\frac{\partial^2 R(Q'; Q)}{\partial q_{\alpha\beta} \partial q'_{\mu\nu}} = \frac{1}{q'_{\mu\nu}} \frac{\partial}{\partial q_{\alpha\beta}} \mathbb{E}_Q[K_{\mu\nu}(t)|y] - \frac{\partial}{\partial q_{\alpha\beta}} \mathbb{E}_Q[S_\mu(t)|y] , \tag{6.1.17}$$

where δ_{ab} is the Kronecker delta. From our previous work, (6.1.16) is easy to obtain, however, (6.1.17) involves derivatives of the expected jumps and holding times and is thus challenging. [BS09] opt for a simple numerical scheme to compute these derivatives and found unstable results, although the authors do remark that more sophisticated numerical schemes could yield improved results at greater computational expense.

We now present the following theorem.

Theorem 6.1.12. *Let $\mu, \nu, \alpha, \beta \in \{1, \dots, h\}$, and Q, Q' be two generator matrices ($\in \Lambda_\epsilon$ for some $\epsilon > 0$). For any two allowed pairs α, β and μ, ν, the derivative in* (6.1.17) *is,*

$$\begin{aligned}
\frac{\partial^2 R(Q'; Q)}{\partial q_{\alpha\beta} \partial q'_{\mu\nu}} = \sum_{s=1}^{n-1} \frac{1}{q'_{\mu\nu}} \Bigg[& -(e^{Q(t_{s+1}-t_s)})^{-2}_{y_s, y_{s+1}} \left(e^{C^{(\alpha\beta)}_\eta (t_{s+1}-t_s)}\right)_{y_s, h+y_{s+1}} (e^{C^{(\mu\nu)}_\gamma (t_{s+1}-t_s)})_{y_s, h+y_{s+1}} \\
& + (e^{Q(t_{s+1}-t_s)})^{-1}_{y_s, y_{s+1}} \left(e^{C^{(\alpha\beta,\mu\nu)}_\psi (t_{s+1}-t_s)}\right)_{y_s, 3h+y_{s+1}} \Bigg] \\
- \Bigg[& -(e^{Q(t_{s+1}-t_s)})^{-2}_{y_s, y_{s+1}} \left(e^{C^{(\alpha\beta)}_\eta (t_{s+1}-t_s)}\right)_{y_s, h+y_{s+1}} (e^{C^{(\mu)}_\phi (t_{s+1}-t_s)})_{y_s, h+y_{s+1}} \\
& + (e^{Q(t_{s+1}-t_s)})^{-1}_{y_s, y_{s+1}} \left(e^{C^{(\alpha\beta,\mu)}_\omega (t_{s+1}-t_s)}\right)_{y_s, 3h+y_{s+1}} \Bigg] ,
\end{aligned}$$

where the $2h$-by-$2h$ matrices, $C^{(\alpha\beta)}_\gamma, C^{(\alpha)}_\phi, C^{(\alpha\beta)}_\eta$, are defined as,

$$C^{(\alpha\beta)}_\gamma = \begin{bmatrix} Q & q_{\alpha\beta} e_\alpha e_\beta^\intercal \\ 0 & Q \end{bmatrix} , \; C^{(\alpha)}_\phi = \begin{bmatrix} Q & e_\alpha e_\alpha^\intercal \\ 0 & Q \end{bmatrix} , \; C^{(\alpha\beta)}_\eta = \begin{bmatrix} Q & e_\alpha e_\beta^\intercal - e_\alpha e_\alpha^\intercal \\ 0 & Q \end{bmatrix} ,$$

and the $4h$-by-$4h$ matrices $C^{(\alpha\beta,\mu\nu)}_\psi, C^{(\alpha\beta,\mu)}_\omega$ are defined

$$C^{(\alpha\beta,\mu\nu)}_\psi = \begin{bmatrix} C^{(\mu\nu)}_\gamma & \frac{\partial C^{(\mu\nu)}_\gamma}{\partial q_{\alpha\beta}} \\ 0 & C^{(\mu\nu)}_\gamma \end{bmatrix} , \; C^{(\alpha\beta,\mu)}_\omega = \begin{bmatrix} C^{(\mu)}_\phi & \frac{\partial C^{(\mu)}_\phi}{\partial q_{\alpha\beta}} \\ 0 & C^{(\mu)}_\phi \end{bmatrix} .$$

The proof of this uses similar techniques to Proposition 6.1.2 along with differentiation properties of matrix-exponentials, and is deferred to Section 6.5.2.

Remark 6.1.13. *In the above representation for the derivative of R, we use subscripts of the form $h + y_{s+1}$ and $3h + y_{s+1}$, this is simply a consequence of the result in* [VL78]*. Namely, we are not interested in all the entries of the matrix, only an h-by-h segment. We therefore need to adjust the indexing to only take elements at this specific segment.*

Using Theorem 6.1.12 and Lemma 6.1.11, we can write the elements of the Hessian corresponding to the $q_{\alpha\beta}q_{\mu\nu}$ derivative as,

$$\begin{aligned}\frac{\partial^2 L(Q;y)}{\partial q_{\alpha\beta}\partial q_{\mu\nu}} =& \sum_{s=1}^{n-1} \frac{-1}{q_{\mu\nu}^2} (e^{Q(t_{s+1}-t_s)})^{-1}_{y_s,y_{s+1}} (e^{C_\gamma^{(\mu\nu)}(t_{s+1}-t_s)})_{y_s,h+y_{s+1}} \delta_{\alpha\mu}\delta_{\beta\nu} \\ &+ \frac{1}{q_{\mu\nu}} \Bigg[-(e^{Q(t_{s+1}-t_s)})^{-2}_{y_s,y_{s+1}} \left(e^{C_\eta^{(\alpha\beta)}(t_{s+1}-t_s)}\right)_{y_s,h+y_{s+1}} (e^{C_\gamma^{(\mu\nu)}(t_{s+1}-t_s)})_{y_s,h+y_{s+1}} \\ &\qquad + (e^{Q(t_{s+1}-t_s)})^{-1}_{y_s,y_{s+1}} \left(e^{C_\psi^{(\alpha\beta,\mu\nu)}(t_{s+1}-t_s)}\right)_{y_s,3h+y_{s+1}} \Bigg] \\ &- \Bigg[-(e^{Q(t_{s+1}-t_s)})^{-2}_{y_s,y_{s+1}} \left(e^{C_\eta^{(\alpha\beta)}(t_{s+1}-t_s)}\right)_{y_s,h+y_{s+1}} (e^{C_\phi^{(\mu)}(t_{s+1}-t_s)})_{y_s,h+y_{s+1}} \\ &\qquad + (e^{Q(t_{s+1}-t_s)})^{-1}_{y_s,y_{s+1}} \left(e^{C_\omega^{(\alpha\beta,\mu)}(t_{s+1}-t_s)}\right)_{y_s,3h+y_{s+1}} \Bigg].\end{aligned}$$

A similar transform to (6.1.12) can be applied here to obtain the Hessian from TPMs. When using the result to estimate the error, some knowledge of the number of companies per rating is required.

Direct Differentiation for Gradient and Hessian of the Likelihood

Relying on first principles, it turns out that one can do without the said formula in [Oak99] and derive a closed form solution involving matrix exponentials for the gradient and the Hessian by direct differentiation.

A formula for obtaining the Hessian is useful, however, while the second derivative can inform us about errors at the level of the generator matrix, it does not shed light on how these errors propagate to the transition probabilities. For that we need to be able to take further derivatives.

Using properties of derivatives and integrals of exponentials of matrices (see [Wil67] and [VL78]) it follows tha for $A \in \mathbb{R}^{h\times h}$,

$$\frac{\partial \exp(At)_{ij}}{\partial a_{\alpha\beta}} = e_i^\intercal \int_0^t \exp(Av)\frac{\partial A}{\partial a_{\alpha\beta}} \exp\big(A(t-v)\big)\mathrm{d}v e_j = e_i^\intercal \exp\left(\left[\begin{smallmatrix} A & \frac{\partial A}{\partial a_{\alpha\beta}} \\ 0 & A \end{smallmatrix}\right] t\right)_{1:h,h+1:2h} e_j. \tag{6.1.18}$$

Using (6.1.18), we can directly calculate the first and second derivative of the likelihood function for a discretely observed Markov process. Let (α,β) and (μ,ν) be allowed pairs for generator Q, then the gradient and Hessian of the logarithm of (6.1.2) are given by. **Gradient,**

$$\frac{\partial \log L(Q|N)}{\partial q_{\alpha\beta}} = \sum_{u=1}^{N}\sum_{i=1}^{h}\sum_{j=1}^{h} N_{ij}(u)\frac{\exp(C_\eta^{(\alpha\beta)}\Delta t_u)_{i,h+j}}{\exp(Q\Delta t_u)_{i,j}} \quad \text{with} \quad C_\eta^{(\alpha\beta)} = \begin{bmatrix} Q & e_\alpha e_\beta^\intercal - e_\alpha e_\alpha^\intercal \\ 0 & Q \end{bmatrix},$$

while for the **Hessian** we have

$$\begin{aligned}H(Q)_{\alpha\beta,\mu\nu} &= \frac{\partial^2 \log L(Q|N)}{\partial q_{\alpha\beta}\partial q_{\mu\nu}} \\ &= \sum_{u=1}^{n}\sum_{i=1}^{h}\sum_{j=1}^{h} \frac{N_{ij}(u)}{\exp(Q\Delta t_u)_{ij}} \left[\frac{\exp(C_\eta^{(\alpha\beta)}\Delta t_u)_{i,h+j}\exp(C_\eta^{(\mu\nu)}\Delta t_u)_{i,h+j}}{\exp(Q\Delta t_u)_{ij}} - \exp(C_\xi^{(\alpha\beta,\mu\nu)}\Delta t_u)_{i,3h+j} \right], \\ &\qquad \text{whereas } C_\xi^{\alpha\beta,\mu\nu} = \begin{bmatrix} C_\eta^{(\alpha\beta)} & \frac{\partial C_\eta^{(\alpha\beta)}}{\partial q_{\mu\nu}} \\ 0 & C_\eta^{(\alpha\beta)} \end{bmatrix}.\end{aligned}$$

These estimates are direct applications of (6.1.18) (see also Section 6.5.2), hence we omit the steps. Both approaches yield are exact expressions for the Hessian and thus for the Fisher information matrix. However, the direct approach is of distinctly reduced complexity, which consequently leads to clearly reduced computing times.

Computation of the Error Since the Hessian is only defined for allowed pairs the matrix is smaller than $(h-1)^2$-*by*-$(h-1)^2$. We compute the Wald confidence intervals as follows,

- Let N_a be the number of allowed pairs in the estimated Q. Define an N_a-*by*-2 matrix V_Q as the matrix which records the allowed pairs of Q. The ijth component of the Hessian is the differential
$$\frac{\partial^2}{\partial q_{V_Q(i,1)V_Q(i,2)}\partial q_{V_Q(j,1)V_Q(j,2)}}\,.$$
- The information matrix is given by $-H(\cdot)$. The estimated variance of the allowed parameter q_{ab} is then the i^{th} diagonal element of $-H(\cdot)^{-1}$, where $V_Q(i,1)=a$ and $V_Q(i,2)=b$.
- The Wald 95% confidence interval for q_{ab} is $q_{ab}\pm 1.96\sqrt{Var(q_{ab})}$.

6.1.3 The Delta method - Confidence Intervals for Probabilities

The object we are estimating is the generator matrix Q, thus the confidence intervals are based on the entries of this matrix. Although it is useful to know the confidence interval for such an estimation, from a practitioners standpoint it is more useful to know how this uncertainty propagates to the underlying TPM and estimated probabilities of default. This is a classical problem in statistics where one wishes to consider how the confidence interval changes under some transformation (in this case $P(t)=\exp(Qt)$), the method to do this is known as the *Delta method*, see [LC98] for further information.

We construct confidence intervals for each element in P individually using the set of *allowed pairs*, see Definition 6.1.10, we consider the confidence interval for the transition probability p_{ij} at time t as,

$$p_{ij}(V_Q;t) := \left(e^{Qt}\right)_{ij}\,.$$

That is for a fixed t, $p_{ij}(V_Q;t)$ is a multivariate function of the allowed pairs, V_Q, in Q. This leads to the following result.

Theorem 6.1.14. *Assume asymptotic normality holds for all allowed pairs, let $V_{\hat{Q}}$ denote the allowed pairs of $\hat{Q}$ (our MLE estimate) and fix t. Then, for each i,j in the state space with $i\neq h$, the variance in p_{ij} is given by,*

$$Var\Big(p_{ij}(V_{\hat{Q}};t)\Big) \approx \frac{\partial p_{ij}(V_{\hat{Q}};t)}{\partial V_{\hat{Q}}}\left(-H(\hat{Q})^{-1}\right)\left(\frac{\partial p_{ij}(V_{\hat{Q}};t)}{\partial V_{\hat{Q}}}\right)^{\intercal}, \tag{6.1.19}$$

provided $\partial p_{ij}(V_{\hat{Q}};t)/\partial V_{\hat{Q}}\neq 0$, where $\frac{\partial}{\partial V_{\hat{Q}}}$ denotes the vector constructed by differentiating w.r.t. each element in $V_{\hat{Q}}$ then evaluated at $\hat{Q}$, and $H(\hat{Q})^{-1}$ is the inverse Hessian matrix at the MLE. Moreover, for each $(\alpha,\beta)\in V_{\hat{Q}}$,

$$\frac{\partial p_{ij}(V_{\hat{Q}};t)}{\partial q_{\alpha\beta}} = \Big(\exp(C_\eta^{(\alpha\beta)}t)\Big)_{i,h+j} \quad \textit{where} \quad C_\eta^{(\alpha\beta)} = \begin{bmatrix} \hat{Q} & e_\alpha e_\beta^\intercal - e_\alpha e_\alpha^\intercal \\ 0 & \hat{Q} \end{bmatrix}.$$

The proof of this result is given in Section 6.5.3. The assumption that $\partial p_{ij}(V_{\hat{Q}};t)/\partial V_{\hat{Q}}\neq 0$ is extremely mild and can be easily checked once the MLE estimate is found.

6.1.4 Further Discussion on the Results

The main results presented throughout this section have been involved with the convergence and error estimation of the EM algorithm. Although we have concentrated on the particular problem of credit risk modelling, the problem of determining transition matrices over shorter time frames is not unique to finance. For example [CS02] consider a similar problem but in a medical setting. The assumptions we make here are appropriate for the finance problem we are considering, however one could look to relax the off diagonal assumption (Assumption 6.1.5) we make on the generator to prove convergence and instead only consider that all ratings (other than default) formed a communicating class. That being said for this particular set up it seems an unnecessary complication. One assumption that is difficult to remove however, is the diagonal dominance assumption required to have a unique (identifiable) generator. That is we need to observe the empirical transition matrix frequently enough so that most transitions are captured. In the majority of cases though I believe this will hold.

Focussing more on the error estimation results and in particular the Delta method, it is not ideal that we must make assumptions that one has to verify a posteriori, such as the derivative of the transition probability being non-zero. However, due to the off diagonal assumption it is not a restrictive assumption, moreover it is also simple to check.

6.2 Competitor Algorithms

We have described several aspects of the EM with relation to this problem. Let us now discuss other algorithms that have been proposed in the literature.

6.2.1 Deterministic algorithms

Diagonal Adjustment (DA) The first method to discuss is diagonal adjustment, see [IRW01]. Given a TPM, P, one calculates the matrix logarithm directly. However, due to the embeddability problem, the logarithm may not be a valid generator. To solve this problem [IRW01] suggest setting for $i \neq j$,

$$q_{ij}^{DA} = \begin{cases} (\log(P))_{ij}\,, & \text{if } (\log(P))_{ij} \geq 0\,, \\ 0\,, & \text{otherwise.} \end{cases}$$

and adjusting (re-balancing) the diagonal element correspondingly, $q_{ii}^{DA} = \sum_{j\neq i} -q_{ij}$ for $i \in \{1, \cdots, h\}$. Hence forcing the corresponding matrix Q^{DA} to satisfy the properties of a generator.

Weighted Adjustment (WA) Weighted adjustment is also suggested in [IRW01]. It follows diagonal adjustment except, one re-balances across the entire row. Again, calculate the logarithm of the TPM to find q's, then compute

$$G_i = |q_{ii}| + \sum_{j\neq i} \max(q_{ij}, 0)\,, \quad B_i = \sum_{j\neq i} \max(-q_{ij}, 0)\,.$$

The entries corresponding to weighted adjustment are defined as,

$$q_{ij}^{WA} = \begin{cases} 0 & \text{if } i \neq j \text{ and } q_{ij} < 0\,, \\ q_{ij} - B_i|q_{ij}|/G_i & \text{otherwise if } G_i > 0\,, \\ q_{ij} & \text{otherwise if } G_i = 0\,. \end{cases}$$

Quasi-Optimisation of the Generator (QOG) The above two methods are unfortunately not optimal in any sense. The QOG (Quasi-Optimisation of the Generator), method suggested in [KS01] relies on optimisation and is therefore an improvement on the diagonal and weighted adjustment methods. QOG involve solves the minimisation problem $\min_{Q\in\mathcal{Q}} \|Q-\log(P)\|$, where $\mathcal{Q}$ is the space of stable generator matrices and $||\cdot||$ is the Euclidean norm. Further, [KS01] provide an efficient algorithm to obtain Q.

6.2.2 Statistical algorithm: Markov Chain Monte Carlo

An alternative statistical algorithm one can adopt is MCMC (Markov Chain Monte Carlo), see Section 5.2. It should be noted that all MCMC algorithms presented here use a so-called auxiliary variable technique, by introducing the fully observed Markov chain, X as a random variable.

Moreover, the prior for Q is $\Gamma(\alpha, 1/\beta)$ (shape and scale), which is conjugate for the likelihood of a CTMC.

Gibbs Sampling - Bladt & Sorensen 2005 To simulate the Markov process, X, [BS05] suggest a rejection sampling method. As is stated in [BS05], such a sampling method runs into difficulties when considering low probability events since the rejection rate will be high (e.g. default for high rated bonds). The MCMC algorithm is summarised as follows, [Ina06],

1. Construct an initial generator Q using the prior distribution ($\Gamma(\alpha_{ij}, 1/\beta_i)$).

2. For some specified number of runs

   (1) Simulate X for each observation from Y, with law according to Q.

   (2) Calculate the quantities of interest K and S, from X.

   (3) Construct a new Q by drawing samples from $\Gamma(K_{ij}(t) + \alpha_{ij}, 1/(S_i(t) + \beta_i))$.

   (4) Save this Q and use it in the next simulation.

3. From the list of Qs, drop some proportion (burn in), then take the mean of the remainder.

The issues with this method are the choice of α and β and the number of runs required before we know that the sample has converged (burn in). Both of these are critical in obtaining accurate answers from MCMC and although [BS05] suggested taking α_{ij} and β_i to be 1, they observe MCMC overestimating entries in the generator when true entries were small. Furthermore, here we are required to use the TPM indirectly through inferring company transitions. That is, we consider M companies in each rating and define the number of companies to make the transition i to j as $M \times P_{ij}$, this of course need not be an integer, but we can always normalise the entries. The reason we cannot use the TPM directly as we did in the EM algorithm is due to the fact that MCMC becomes very sensitive to the values in the prior. The burn in for MCMC will be of little concern to us here as will become apparent when carrying out analysis on the algorithms.

Importance Sampling - Bladt & Sorensen 2009 [BS09] address some of the issues in [BS05] by running the same algorithm as previous combined with an importance sampling scheme based on the Metropolis-Hastings algorithm (in its essence a single component Metropolis-Hastings algorithm). The proposal distribution suggested is a Markov chain with generator given by the 'neutral matrix' Q^*, which takes the following form,

$$Q^* = \frac{1}{W}\big(\mathbf{1}_h - \mathbf{I}_h - h\mathbf{I}_h\big)\,,$$

where $\mathbf{1}_h$ and $\mathbf{I}_h$ is the h-by-h matrix of ones and identity matrix respectively and W is a scaling factor set to match the intensities in the true generator matrix Q. [BS09] note that if entries in Q are known to be zero, then the corresponding element in Q^* should also be set to zero and the diagonal modified accordingly. Thus transitions rarely produced by the generated Markov chain will occur much more frequently under Q^*. Thus we have solved (at least partially) one of the problems faced in MCMC. The importance sampling weights for a chain X are,

$$w(X) = \frac{L(Q;X)}{L(Q^*;X)}\,,$$

where L is the CTMC likelihood. For the priors, [BS09] do not suggest any significant improvement on their earlier work. The authors use $\alpha = 1$ and $\beta = 5$, which they claim gives better results than the suggestion in [BS05]. However, it still provides a problem when dealing with entries in Q which are close to zero. The problem stems from the fact that very little information is known (rarely observed) for certain transitions, therefore the output for these entries is mostly based on our prior beliefs.

MCMC Mode Algorithm [Ina06] presented an alternative algorithm to the original MCMC algorithm presented in [BS05] whereby one calculates the mode rather than the mean. The author claims that this gives extremely accurate results and outperforms other algorithms. The reasoning presented is that the standard MCMC overestimates in the small probability cases due to the gamma distribution being 'skewed', therefore the mode is a better estimate. [Ina06] approximates the mode of $\{q_{ij}^{(k)}\}$ by kernel smoothing over the estimates (after taking the log transform to ensure all results are positive).

6.3 Benchmarking the Algorithms

Due to the diversity of investments a bank makes, one cannot assess an algorithms' performance with a single test. With this in mind we consider a host of tests on different portfolios and matrices. The computations were carried out on a Dell PowerEdge R430 with four Intel Xeon E5-2680 processors. At the same time the work [SdR17], [Pfe17] also carried out a comparative study. The performance tests of [Pfe17] are a just subset of those we present next and independently confirm (where there is overlap) our findings, in particular the timing of the MCMC algorithms versus the EM. A version of our algorithms appear in the mentioned R-package (see Remark 6.0.2).

The tests conducted in this section are theoretical in the sense that we take a "true" solution (based on reasonable values). This is important for a comparative analysis because it allows us to measure the suitability of the model in terms of convergence etc. Later in this chapter and in Chapter 7 we shall use empirical data to demonstrate the model and techniques.

The first observation we make is, transition matrices can vary substantially depending on the financial climate (see [CHL04] and [Can04]). Therefore we consider two different generator matrices which can be thought of as the generator in financial stress and the generator in financial calm. In order to keep these matrices 'reasonable' we start off with the generator given in [CHL04] built using a large amount of data (see also [Ina06]) and consider a generator which has in general higher transition rates and one with lower transition rates. Through considering more than one generator this provides a more detailed assessment of the performance of the various algorithms than other comparative reviews, such as [Ina06]. The generators we consider are shown in Table 6.1 and Table 6.2. We observe that Table 6.1, has more non-zero entries and larger entries than that of Table 6.2.

	AAA	AA	A	BBB	BB	B	C	D
AAA	-0.146371	0.085881	0.04549	0.015	0	0	0	0
AA	0.018506	-0.166337	0.114831	0.033	0	0	0	0
A	0.0276	0.047012	-0.198043	0.09043	0.023001	0.01	0	0
BBB	0.011469	0.010734	0.088133	-0.243046	0.077569	0.044407	0.010734	0
BB	0	0	0.019159	0.184699	-0.323077	0.106166	0.013053	0
B	0	0	0.012280	0.034822	0.093489	-0.296265	0.134273	0.022401
C	0	0	0	0	0.02	0.140209	-0.600939	0.440730
D	0	0	0	0	0	0	0	0

Table 6.1: True unstable generator

	AAA	AA	A	BBB	BB	B	C	D
AAA	-0.061371	0.055881	0.005490	0	0	0	0	0
AA	0.013506	-0.096337	0.074831	0.008	0	0	0	0
A	0	0.037012	-0.097442	0.06043	0	0	0	0
BBB	0	0.000734	0.058133	-0.120843	0.057569	0.004407	0	0
BB	0	0	0.009159	0.104699	-0.190024	0.076166	0	0
B	0	0	0	0.024822	0.083489	-0.174985	0.064273	0.002401
C	0	0	0	0	0	0.080209	-0.300939	0.220730
D	0	0	0	0	0	0	0	0

Table 6.2: True stable generator

Throughout the analysis we refer to the multiple MCMC algorithms introduced in Section 6.2 which we label in the following way: `MCMC BS05` is [BS05]'s algorithm of Section 6.2.2;

`MCMC BS09` is [BS09]'s algorithm of Section 6.2.2; and `MCMC Mode` is [Ina06]'s algorithm in Section 6.2.2.

6.3.1 Sample Size Inference

The first test we consider is an extension to a test in [Ina06], where the author considers a true underlying generator and masks it by using it to simulate TPMs, which we view as observations, then applying the algorithms to each observation. The key point here is, [Ina06] only simulates 100 companies per rating and hence the outputted TPM is non-embeddable (has 0 entries for accessible jumps). This is an extremely useful test because it provides a fair and intuitive way to assess the performance of each algorithm, however, [Ina06] only considers one true generator and only one level of information i.e. 100 companies per rating. Alongside the two different generators we also consider a range of companies per rating to determine its effect on convergence for each algorithm. Furthermore, [Ina06] uses seven years worth of data, although one would likely have access to multiple years worth of TPM data, it is unlikely that we would have seven years of transitions from the same generator. Hence we consider four years, which is more consistent with time homogeneity estimates for generators (see [CHL04]). We calculate our estimates for the generator as follows.

1. Take a range of obligors per rating, $[100, 200, 300, 500, 750, 1000]$ and 10 random seeds.
2. For each true generator simulate four one year TPMs for each seed and for each obligor per rating. Hence we have (#Years×#Obligors categories×#Random Seeds×#True generators), simulated TPMs.
3. For each set of four simulated TPM we estimate the generator for each algorithm. MCMC may take a long time to run, therefore we consider the time taken to carry out the first 10 runs and the total time taken, if these exceed 180 or 18 000 seconds respectively, the algorithm is deemed to be too slow and no result is returned. Note, MCMC algorithms use 3000 runs with a burn in of 300. This is smaller than [Ina06] for example, however, [Ina06] shows apparent convergence to the stationary distribution in a small number of iterations and we observe a similar result.
4. Therefore, for each algorithm we have (# Obligors categories × # Random Seeds × # True generators) estimated generators to analyze.

We analyze the estimated generators by considering, distance between estimated generator and true generator in Euclidean norm and difference in one year probability of default. All results presented have been obtained by analyzing the estimated generator for each seed, then averaging. This gives a better picture of the average performance.

Algorithms	Deterministic	EM	MCMC
Time (seconds)	< 1	~ 10	$\sim 10^3$ to $\sim 10^4$

Table 6.3: Order of time taken to execute the various algorithms. Note that MCMC also depends on the level of information i.e. obligors in each rating. We also note that BS 09 algorithm is faster than the other MCMC algorithms but still takes 10^4 seconds in the case of 1000 obligors per rating.

Convergence in Euclidean Norm Our goal in this analysis is to consider the empirical rate of improvement of each algorithm as our 'information' about the true generator increases. For each obligor category we calculate the natural log of the distance (measured by the Euclidean norm) between the estimate and the true. The results are shown in figures 6.1 and 6.2.

Note the x-axis is on a logarithmic scale. We observe similarities between the two figures, most notably in the case of low information all algorithms have very similar convergence results, however, as we increase the information there is substantial variation in improvement, `MCMC BS09` algorithm does not improve as well as the other algorithms. Missing points stem from an algorithm failing the acceptance times.

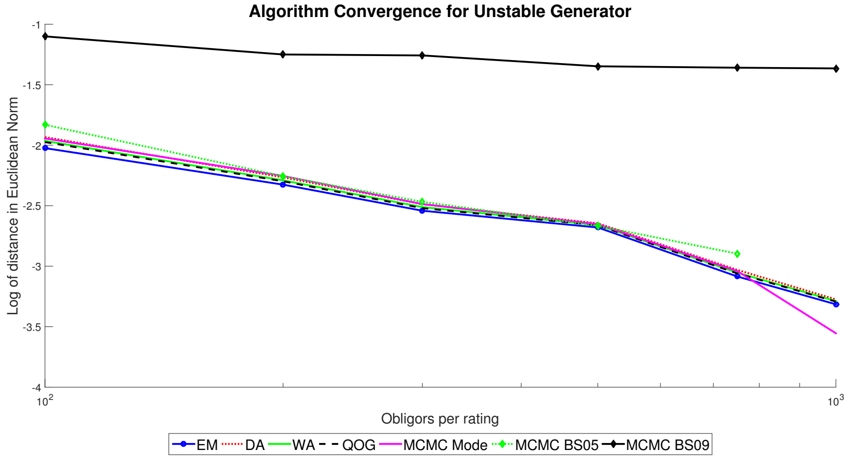


Figure 6.1: Showing the log of the error for each algorithm as a function of obligors per rating.

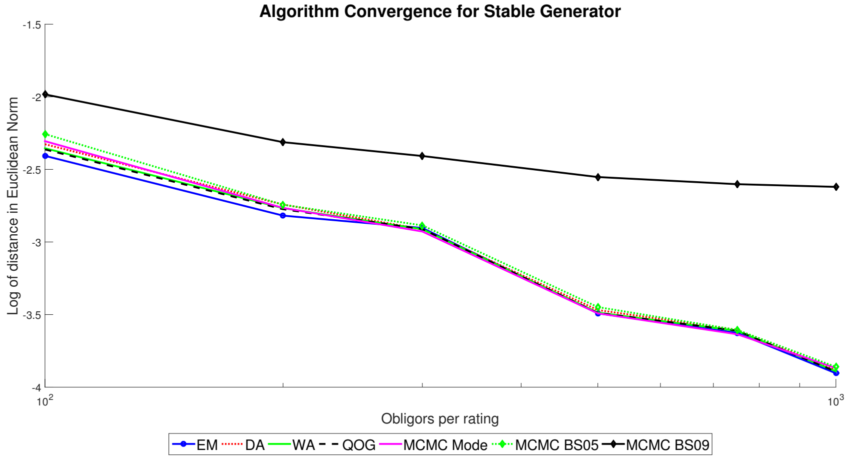


Figure 6.2: Showing the log of the error for each algorithm as a function of obligors per rating.

The MCMC algorithms have a potentially increased error due to the Monte Carlo simulation, lowering it requires a larger computational expense to the already most expensive algorithm being tested here. For the [BS09] algorithm, the neutral matrix approximation may give poor mixing, thus the additional error.

Error in Probability of Default Although overall error is important, it does not provide details on the small probability scale. This is extremely important in banking, since estimation of the probability of default is crucial. Using the same estimated generators as previous we calculate the corresponding one year TPM, that is, we calculate $\exp(Q_{\text{estimate}})$ (using the `expm` function in MATLAB) for each seed then take the average. The averaged TPM default probabilities are compared to the true ones. To keep the numbers in the comparisons meaningful we plot the log of the relative error, where we define,

$$\text{Relative Error} = \frac{|\text{PD}_{\text{estimate}} - \text{PD}_{\text{true}}|}{\text{PD}_{\text{true}}} .$$

The results of which are given in Figures 6.3 and 6.4.

Unlike the overall error, there appears to be far greater volatility in the error estimation w.r.t. the probability of default. Moreover, there appears to be no general downward trend in error for the investment grade ratings. A likely cause for this is, even with 1000 companies there are still no/few investment grade defaults. Of the algorithms `MCMC BS09` performs the worst. The EM algorithm though has consistently one of the smallest errors and is clearly the best in the investment grades.We have only shown the results for the unstable generator, the stable

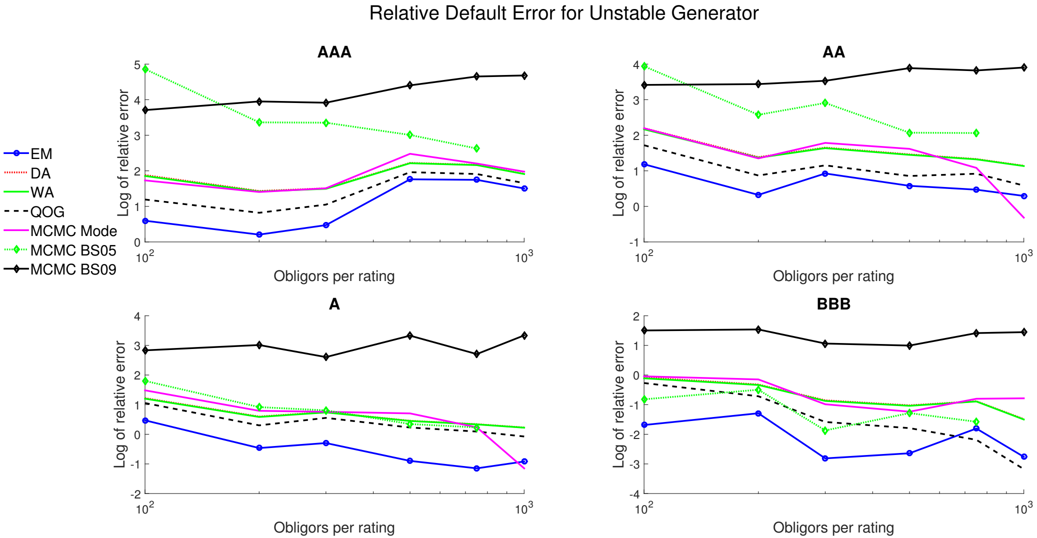


Figure 6.3: Showing the log of the relative default error for each algorithm as a function of obligors per rating.

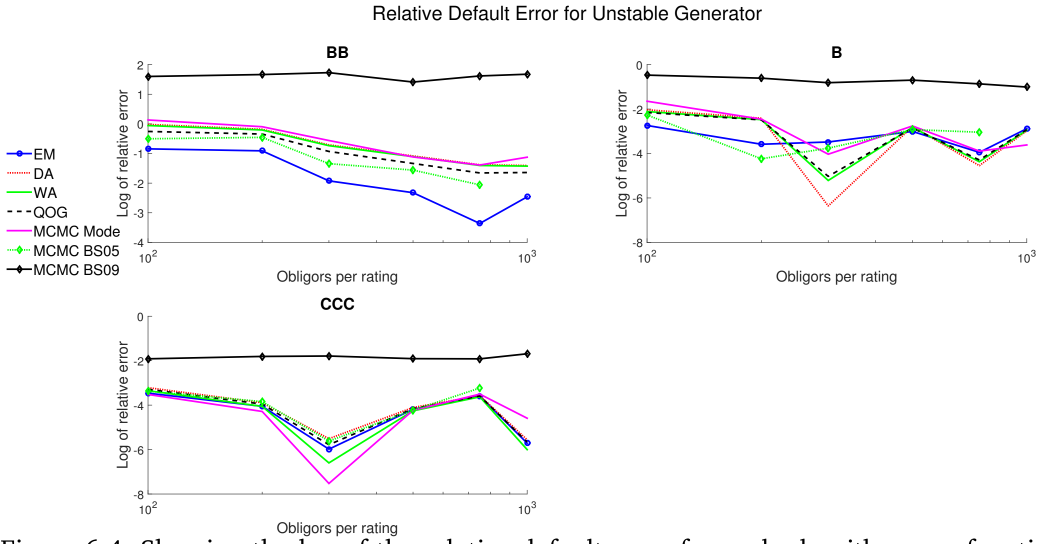


Figure 6.4: Showing the log of the relative default error for each algorithm as a function of obligors per rating.

generator was similar.

6.3.2 Time Dependent Probability of Default

A key question that has not been addressed in the literature is how do the probabilities of default change in time among the several algorithms. For ths we only consider EM, QOG, WA and the MCMC Mode algorithm from [Ina06], since these algorithms gave the best probability of default estimates.

We consider a non-embeddable TPM, then estimate the generator matrix Q, from Q we can easily calculate the probability of a company with some initial rating defaulting in time $t > 0$. The goal here is to assess how that probability changes with time. The TPM is given in Table 6.4, for the MCMC algorithm we took this table to be generated with 250 obligors per rating.

	AAA	AA	A	BBB	BB	B	C	D
AAA	0.8824	0.1176	0	0	0	0	0	0
AA	0.0064	0.9111	0.0813	0.0008	0.0001	0	0.0003	0
A	0.0003	0.0559	0.8836	0.0499	0.0079	0.0015	0.0002	0.0007
BBB	0	0.0116	0.1585	0.7640	0.0528	0.0070	0	0.0061
BB	0	0	0.0213	0.1193	0.7746	0.0623	0.0099	0.0127
B	0	0	0.0062	0.0199	0.1669	0.7017	0.0730	0.0322
C	0	0	0	0	0.0417	0.2083	0.4544	0.2956
D	0	0	0	0	0	0	0	1

Table 6.4: Observed TPM used to estimate the generators in probability of default plots.

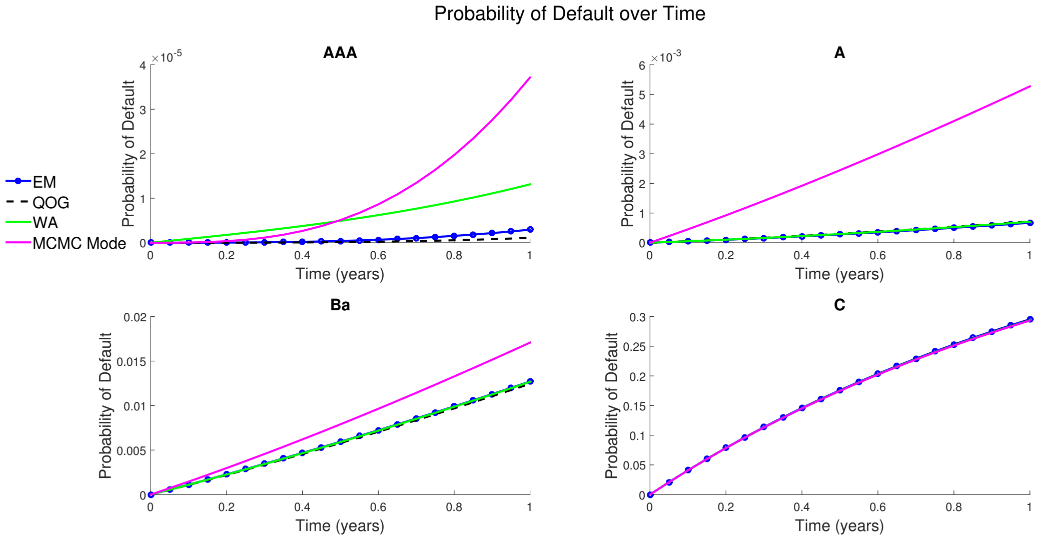


Figure 6.5: Probability of default over time for EM, QOG, MCMC Mode and WA.

The probability of default across ratings over the one year time horizon is found in Figure 6.5. The plots give a deeper understanding to the algorithms themselves. As the probability of default increases the algorithms converge, however, in the case of less defaults we observe a much larger discrepancy. This can be thought of as the algorithm's ability to deal with missing data, in the lower grades we observe defaults and thus have a handle on the probability, however, in the case of AAA ratings we observe no defaults and therefore it is an approximation by the algorithm. This shows the difference between the methods, shows the potential prior dependence in the MCMC algorithm. What is also extremely interesting is that QOG set the jump in the generator from AA to C as zero (even though the TPM has a non zero entry there), this implies QOG may in some places under estimate the risk for the investment ratings, this can be seen by the fact QOG puts a smaller probability of default on AAA.

There is a clear ovestimation of the probability of default at higher grades by the WA and MCMC algorithms.

6.3.3 Risk Charge

The previous tests have been rather theoretical, we now consider a practical test to asses the performance of these algorithms in calculating risk charges. We do not give much discussion to the calculation of these risk charges for more technical details readers should consult texts such as [SC11]. Here we consider multiple stylized portfolios to represent the risk appetites of different banks. To best of our knowledge analysis into how different risk measures react to different portfolio types has not been considered in the literature. The risk charges we consider are IRC (VaR at 99.9% with a 3 months liquidity horizon including mark to market loss), IDR (VaR at 99.9% over one year only considering default) and a theoretical risk charge which is IRC but measured using Expected Shortfall (ES) at 97.5%. The final risk charge is included due to the Basel committee showing an increasing interest in ES. We consider 4 years worth of simulated data, and to keep the analysis realistic we consider 200 companies per rating. We consider 3 different portfolios corresponding to risk averse (all investment grade), a speculative portfolio (all speculative grades) and finally a mixed portfolio. The portfolios considered are given in Tables 6.5, 6.6 and 6.7. The tables show the values and ratings of the various bonds in each portfolio.

AAA	100, 500, 1500, 750
AA	200, 750, 2000, 650
A	150, 400, 400
BBB	300, 500, 150, 1500
BB	500, 250, 700
B	200, 500
C	100, 150, 200

Table 6.5: Mixed portfolio

AAA	1000, 500, 1500, 1500
AA	100, 400, 750, 2000, 400, 1500
A	150, 100, 800, 400, 200
BBB	
BB	
B	
C	

Table 6.6: Investment portfolio

AAA	
AA	
A	
BBB	
BB	1000, 150, 100, 800, 1500
B	100, 300, 400, 750, 2000, 1500
C	400, 500, 400, 1000

Table 6.7: Speculative portfolio

Alongside these portfolios we calculate the risk charges using the following information,

- The interest rates we receive for a bond in each rating are

AAA	AA	A	BBB	BB	B	C
2.65%	2.69%	2.78%	2.93%	3.18%	5.45%	12.39%

  These figures are based on interest rates from Moody's and can be found in Section 4.1 of [SC11]. Although these interest rates do not technically match the generators we are using for the TPMs they provide reasonable interest rates for our toy example.

- We assume that all money is lost in the case of default (zero recovery rate).

- We calculate credit migration using the one factor[¶] credit metrics model ([GFB97]), i.e. normalised asset returns follow,

$$z_i = \beta_i X + \sqrt{1 - \beta_i^2}\epsilon_i \, ,$$

  where X is the systematic risk, ϵ_i is the idiosyncratic risk both standard normally distributed and β_i is the correlation to the systematic risk, defined in [Sup03, p.50],

$$\beta_i = 0.12 \left(\frac{1 - \exp\{-50 \, P_i^D\}}{1 - \exp\{-50\}} \right) + 0.24 \left(1 - \frac{1 - \exp\{-50 \, P_i^D\}}{1 - \exp\{-50\}} \right) \, ,$$

  where P_i^D is the probability of default of asset i. Consequently we see that the higher P_i^D the lower the value of β.

- Although more sophisticated methods are available for calculation of VaR and ES (see [Fer14]), we calculate the risk charges using Monte Carlo. This is sufficient here since the portfolios are small relative to a typical bank portfolio, therefore we can obtain accurate estimates using a reasonable number of simulations.

- Again, we calculate 10 realisations of the TPMs and estimate a generator for each.

We consider 15$\times 10^5$ simulations for each portfolio, to assess whether this was sufficient we calculated VaR and ES using 7.5$\times 10^5$, 10$\times 10^5$, 12.5$\times 10^5$ and 15$\times 10^5$ simulations and found the difference between 7.5$\times 10^5$ and 15$\times 10^5$ to be $< 5\%$ for all cases. Hence were are confident that 15$\times 10^5$ gives sufficiently accurate results for our purposes.

With respect to the risk charge calculation, similar to the previous analysis, we calculate the risk charges for every set of TPMs, then average over all the seeds to obtain the risk charge. The risk charges as set by the true generators are given in Table 6.8.

[¶] This is technically not the true regulation for the calculation of IDR which requires a two factor model, however our goal here is only to use these calculations as a method for comparing algorithms.

	Stable			Unstable		
	Mixed	Investment	Speculative	Mixed	Investment	Speculative
IRC	702	0.32	3395	1251	0.41	5057
IRC ES	508	0.20	2409	842	3.78	3826
IDR	750	0	3400	1750	200	4600

Table 6.8: Risk charge results for the true generators.

To asses the performance of each algorithm we measure the error by the following,

$$\text{Risk Error} = \frac{\frac{1}{N}\sum_{i=1}^{N} |\text{Risk Charge Estimate}(i) - \text{Risk Charge True}|}{\text{Risk Charge True}},$$

where Risk Charge Estimate(i) is the i^{th} realisation of the risk charge and N is the number of TPM sets (10 here). The results obtained by the algorithms are shown in Table 6.9.

		Stable			Unstable		
		Mixed	Investment	Speculative	Mixed	Investment	Speculative
IRC	EM	7.3	7.5	1.5	22.5	29 195	2.6
	DA	11.9	8.1	2.4	36.9	66 829	4.3
	WA	11.8	8.1	2.3	37.3	69 293	4.1
	QOG	11.6	7.8	2.3	26.7	38 976	4.1
	MCMCBS05	154	306 000	2	49.6	478 000	4.1
	MCMCBS09	24.9	18.4	14.4	68.3	264 000	14
	MCMCMode	12.5	8.1	3.6	34.9	39 000	3.9
IRC ES	EM	5.3	115	3.4	8.6	375	2.7
	DA	8.2	235	5.1	16.6	1130	3.9
	WA	7.8	210	5	16.4	1109	3.8
	QOG	7.3	123	4.9	12.6	622	3.8
	MCMCBS05	35.4	135 000	4.7	19.7	5315	4.1
	MCMCBS09	21	610	15.5	67.7	6693	13.1
	MCMCMode	9.2	235	6.1	19.1	1063	3.5
IDR	EM	6	0	0.3	4.3	113	3.5
	DA	10	0	1.2	8.6	295	5.7
	WA	9.3	0	0.6	8.6	295	5.2
	QOG	7.3	0	0.6	5.4	185	5.3
	MCMCBS05	139	1580	0.3	12.6	530	4.7
	MCMCBS09	20	40	9.3	33.7	775	13.2
	MCMCMode	10	10	0.9	8	278	4.7

Table 6.9: Risk charge results for each algorithm as a %.

It should be noted, in the stable IDR some algorithms produce a non-zero value for the investment portfolio, therefore we have inserted the money value. The first observation we make is, all algorithms overestimate the risk for the investment portfolio. This is down to two key feature, one is the 'step like' nature of VaR, where in a small portfolio, small probability changes can make a large difference. The other is because we are averaging over multiple Monte Carlo simulations, thus having one default in one of those realisations will change the overall average dramatically. In terms of a typical bank portfolio this type of error should not be a problem since we would be dealing with a far larger number of assets and hence one would obtain multiple defaults. However, the results do still give a useful comparison between the algorithms. Although the MCMC algorithms can outperform the deterministic algorithms for the speculative grades, remarkably in all categories the EM produces the best results. From the tests we have considered we conclude the EM to be the superior algorithm for this problem.

6.4 Error estimation of the EM algorithm

As we have established that the EM algorithm is a good choice of algorithm for this problem let us now also discuss another important feature of the algorithm, namely the ability to obtain (cheaply) Wald confidence intervals for the estimate. This is of course in general a major advantage of the statistical algorithms over their deterministic counterparts is that one can derive error estimates (confidence intervals) without the brute force (slightly ad-hoc) method of bootstrapping. For MCMC this comes by looking at the posterior distribution, which we get for free. However, as we have seen MCMC is computationally expensive. Moreover, unlike the EM where the error can be transformed cheaply, MCMC requires use to calculate stochastic matrices for each realisation then calculate the credible interval.

For this we use empirical data, namely, we take the full dataset described in Section 7.1 (which uses Moody's ratings) and build the corresponding one year TPM.

Remark 6.4.1 (Difference in the notation)**.** *Our previous examples were based on Standard and Poor notation. However, now that we have access to the Moody's data we use this. The key differences are that there is one extra rating (9 instead of 8) and default is now denoted by C.*

A 95% confidence interval for the generator matrix estimate based on Moody's discretely observed corporate ratings data for the year 2016 is illustrated in Figure 6.6. For this interval, the computation time was 0.8s for the direct differentiation approach compared to 1.9s for the Oakes formula approach, both detailed in 6.1.2. For the 21 dimensional generator matrix confidence interval of Moody's corporate ratings data with the modifiers 1, 2 and 3 (and aggregated annual transitions from 1987 to 2016) the computing times are 35.5s for the new expression vs. 83.9s for the formula of [dRS17].

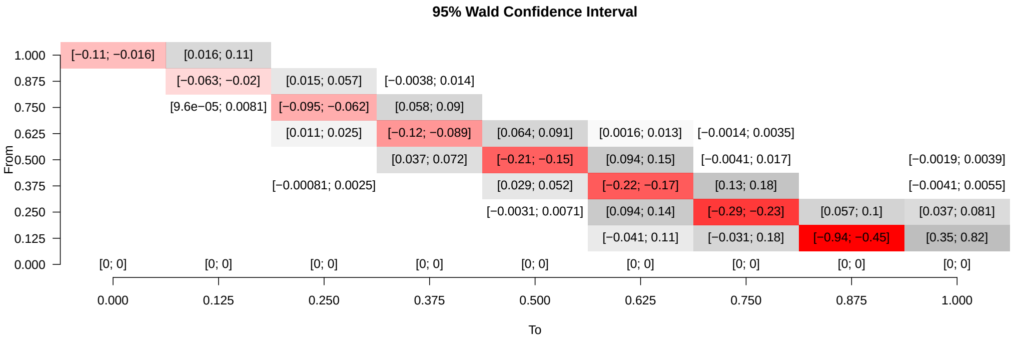


Figure 6.6: Confidence Interval for the entries of the Generator Matrix for Moody's Corporate Rating Discrete-Time Transition Matrix 2016.

As we have already derived a closed form expression for the Hessian we can easily compute (6.1.19). Hence, it is straightforward to compute the confidence interval for the transition probabilities. This is of course an extremely useful result since it allows one to understand the uncertainty at the level of the estimation of transition probabilities, and critically, uncertainties in the probability of default. Figures 6.7 and 6.8 show such intervals for probability of default estimates from Moody's corporate ratings data 2016 and a time horizon of up to 10 years. One can see that this procedure allows to quantify the error of probability of default predictions for arbitrary time horizons. This is especially interesting as this parameter is an important ingredient to the calculation of expected losses over lifetime in the IFRS 9 regulatory framework.

6.4.1 Confidence Intervals w.r.t. information

The previous example looked at confidence intervals with fixed information. Now we want to look at how these intervals change w.r.t. new (more) information. We consider a true generator matrix (which is the MLE Markov generator from the full dataset, described in Section 7.2.5) and from that simulate multiple companies over multiple years to construct a "empirical" dataset.

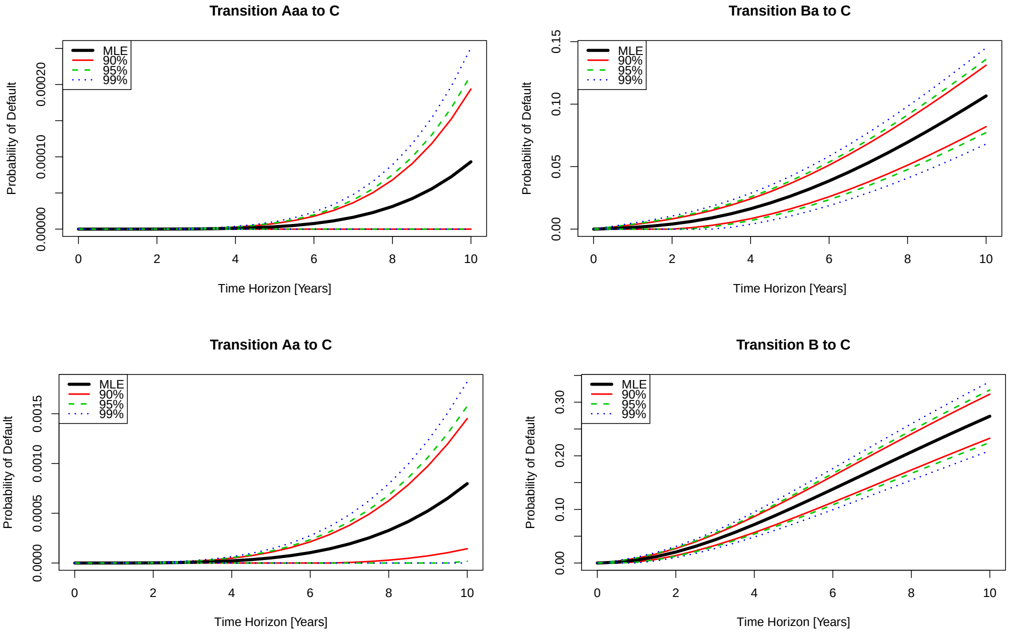


Figure 6.7: Confidence Intervals as maps of time for Discrete-Time Transitions into the Default Category C over 10 years- Moody's Corporate Rating Discrete-Time Transitions 2016

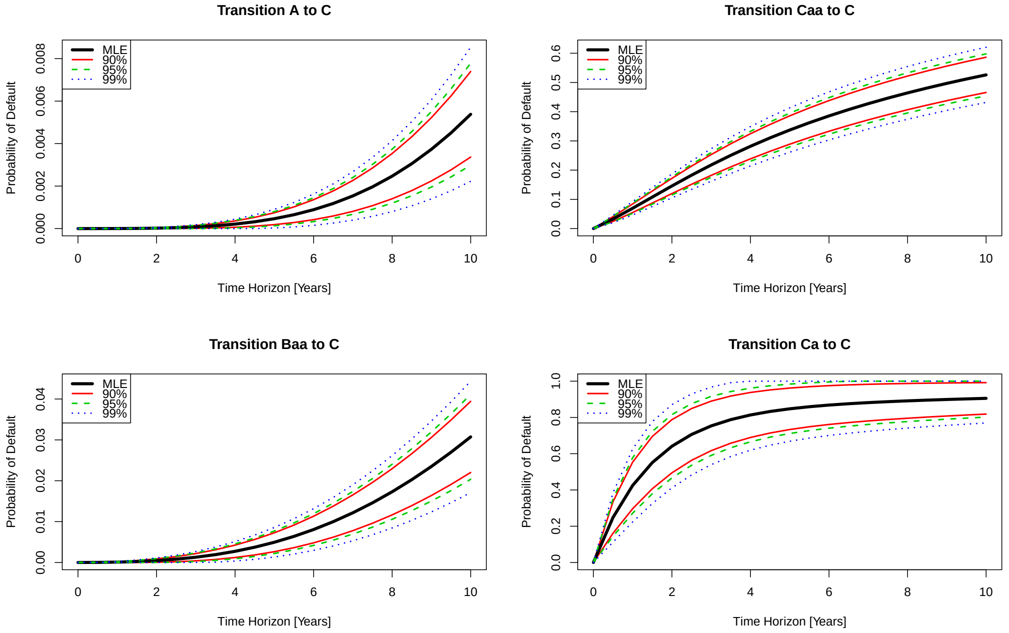


Figure 6.8: Confidence Intervals as maps of time for Discrete-Time Transitions into the Default Category C over 10 years- Moody's Corporate Rating Discrete-Time Transitions 2016

We then introduce the EM algorithm to more and more data and assess how the estimate and error change as the amount of data increases. By using a known generator, we also assess the accuracy of the estimate and error. From a computational point of view, matrix exponentials embed highly nonlinear dependencies in the elements of Q and P therefore to understand the error we consider how both of them change as the amount of information changes.

We consider the scenario of 250 obligors per rating and simulate 50 years worth of transitions (i.e. the number of companies that made each transition). We then apply the EM algorithm using 1 year worth of data then 2 years etc up to 50 years. In the case of a company defaulting we replace it with the rating they were pre-default. This implies that the amount of "information" obtained from each year is similar. We plot the results in Figure 6.9.

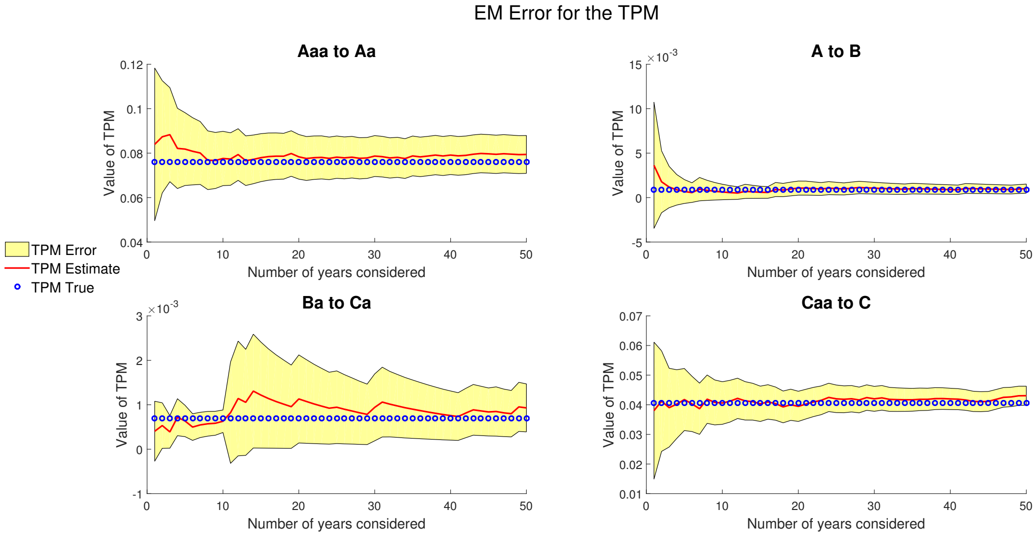


Figure 6.9: Estimated value in some TPM entries and 95% confidence interval as the amount of data increase.

One observes that in most cases the errors in the TPM behave as expected. The surprising result is the Ba to Ca entry, which actually has an increase in error. As alluded above, one can only understand the error in the TPM by understanding the underpinning error of the generator estimation. Although, in theory the Ba to Ca transition depends on all entries in the generator we know that certain entries will have a greater impact. We therefore look at the error in some important generator entries, Figure 6.10.

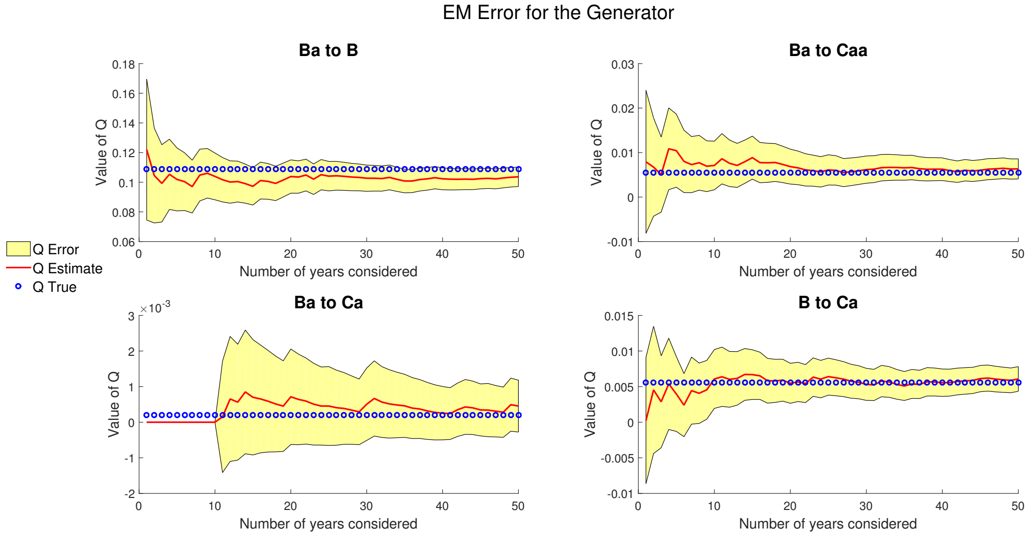


Figure 6.10: Estimated value in the generator and 95% confidence interval as the amount of data increase.

We see that the main contributor to the error is (unsurprisingly) the Ba to Ca entry. Initially we need to wait for a transition from Ba to Ca which increases the likelihood and hence uncertainty surrounding the estimate, moreover it then takes several more years of data before the estimate becomes more stable. This uncertainty in the generator then propagates to uncertainty in the TPM, and one observes the extremely strong correlation between the TPM entry and the corresponding generator entry. Due to this, the error in the Ba to Ca transition probability is much larger than the other estimates, even after 50 years of observation. Of course this behaviour in the CTMC modelling is not ideal (and the IFRS 9 regulation exacerbates the effect), but it shows some of the challenges in obtaining good estimates and errors for small probabilities (rare events), namely that the model is still sensitive to individual observations. One can use this to assess the sensitivity in the model. For example, adding one observation of a company defaulting and recomputing the probabilities with associated error will give us an idea of the sensitivity.

Connection to the Global Maximum

A previous problem with the EM was one could not be sure of the nature of the stationary point. However, we know the form of the Hessian, and therefore we can easily check if this point is a maximum by assessing the eigenvalues of this matrix. Clearly, if we were not at a maximum, then it would be worth perturbing the outputted generator and rerunning the algorithm. As discussed in Remark 6.1.9 the question of a global maximum is very difficult in this setting.

Remark 6.4.2. *One way that has been suggested to improve the chances of the EM converging to the global maximum is, to start from multiple points. Here we can consider creating starting points by setting for each $i \neq j$, $q_{ij} \sim$ Exp(λ) (exponential random variable with intensity λ) for an appropriate λ then setting q_{ii} appropriately.*

We tested the EM according to the above remark and found in every case considered the EM always returns the same generator.

6.5 Proof of Results

6.5.1 Proof of Lemma 6.1.6

We now provide the proof of Lemma 6.1.6, all terms used have the same definition as they did when the Lemma was stated. Throughout we assume $i \neq h$, thus from from Assumption 6.1.5 $\mathbb{P}_Q[X(t) = j|X(0) = i] > 0$ for all $j \in \{1, \dots, h\}$ and $t > 0$. The first inequality we prove is the lower bound on the expected number of jumps. Following the assumptions in Lemma 6.1.6 and time homogeneity we make the observation

$$\mathbb{E}_Q[K_{ij}(T)|P] \geq P^u_{ij}\mathbb{P}_Q[K_{ij}(t) \geq 1|X(0) = i, X(t) = j] \, .$$

The above inequality holds because we are only considering $X(0) = i$, $X(t) = j$ and not all possible combinations of start and end states, moreover, $\mathbb{P}_Q[K_{ij} \geq 1|X(0) = i, X(t) = j] \leq \sum_{n=1}^{\infty} n\mathbb{P}_Q[K_{ij} = n|X(0) = i, X(t) = j]$. We further observe,

$$\mathbb{P}_Q[K_{ij} \geq 1|X(0) = i, X(t) = j] \geq \frac{q_{ij}}{-q_{ii}} \, .$$

Thus the lower bound in inequality (6.1.13) can be easily obtained. We now prove the upper bound on the expected number of jumps. The first observation we make is for all $\nu \in \{1, \dots, h\}$,

$$\mathbb{E}_Q[K_{ij}(T)|X(0) = i, X(t) = \nu] = \sup_{\mu \in \{1,\dots,h\}} \mathbb{E}_Q[K_{ij}(T)|X(0) = \mu, X(t) = \nu] \, .$$

To see this, let $\mu \neq i$, then denote by τ_i the first time the process enters state i (if $\mathbb{P}_Q[X(t) = i|X(0) = \mu] = 0$ for $t > 0$, then the result is trivial), by the law of total probability we find,

$$\begin{aligned}
&\mathbb{E}_Q[K_{ij}(t)|X(0) = \mu, X(t) = \nu] \\
&\qquad = \mathbb{E}_Q[K_{ij}(t)|X(0) = \mu, X(t) = \nu, \tau_i < t]\mathbb{P}_Q[\tau_i < t|X(0) = \mu, X(t) = \nu] \\
&\qquad\qquad + \mathbb{E}_Q[K_{ij}(t)|X(0) = \mu, X(t) = \nu, \tau_i \geq t]\mathbb{P}_Q[\tau_i \geq t|X(0) = \mu, X(t) = \nu] \, .
\end{aligned}$$

The second term is zero. Then, using the Markov property we obtain,

$$\begin{aligned}\mathbb{E}_Q[K_{ij}(t)|X(0)=\mu, X(t)=\nu] &\leq \mathbb{E}_Q[K_{ij}(t)|X(\tau_i)=i, X(t)=\nu, \tau_i < t] \\ &\leq \mathbb{E}_Q[K_{ij}(t)|X(0)=i, X(t)=\nu]\,.\end{aligned}$$

Consequently from this observation and (6.1.12) we obtain,

$$\mathbb{E}_Q[K_{ij}(T)|P] \leq hn \sum_{\nu=1}^{h} \mathbb{E}_Q[K_{ij}(t)|X(0)=i, X(t)=\nu]\,.$$

Observe that,

$$\mathbb{E}_Q[K_{ij}(t)|X(0)=i, X(t)=\nu] = \frac{\mathbb{E}_Q[K_{ij}(t)\mathbb{1}_{\{X(t)=\nu\}}|X(0)=i]}{\mathbb{P}_Q[X(t)=\nu|X(0)=i]} \leq \frac{\mathbb{E}_Q[K_{ij}(t)|X(0)=i]}{\mathbb{P}_Q[X(t)=\nu|X(0)=i]}\,.$$

The numerator is easy to bound by considering the expected number of jumps out of i,

$$\mathbb{E}_Q[K_{ij}(t)|X(0)=i] \leq -q_{ii}t\,.$$

The denominator requires further analysis, firstly, let $m = |i-\nu|$, and therefore by Assumption 6.1.5 we can go from state i to ν in m jumps, w.l.o.g. let $i \geq \nu$ (it will be come clear that the ordering does not matter). Firstly, if $i = \nu$ then,

$$\mathbb{P}_Q[X(t)=\nu|X(0)=i] \geq e^{q_{ii}t}\,.$$

For $i > \nu$, we use the Markov property to obtain,

$$\mathbb{P}_Q[X(t)=\nu|X(0)=i] \geq \prod_{a=1}^{m} \mathbb{P}_Q\left[X\left(\frac{a}{m}t\right) = i+a \middle| X\left(\frac{a-1}{m}t\right) = i+a-1\right]\,.$$

Conditioning on X only making one jump in each increment we obtain,

$$\begin{aligned}\mathbb{P}_Q[X(t)=\nu|X(0)=i] &\geq \prod_{a=1}^{m} \frac{q_{i+a-1,i+a}}{-q_{i+a-1,i+a-1}}(-q_{i+a-1,i+a-1})t\exp(q_{i+a-1,i+a-1}t) \\ &\geq \prod_{a=1}^{m} \epsilon t \exp(-ht/\epsilon)\,.\end{aligned}$$

As $m \leq h$ and the terms are strictly smaller than 1, the sought result follows (independent of $\nu \neq i$).

The last inequality to prove concerns the holding times. By taking for $P^u_{ii} > 0$,

$$\mathbb{E}_Q[S_i(T)|P] \geq P^u_{ii}\mathbb{E}_Q[S_i(t)|X(0)=i, X(t)=i] \geq P^u_{ii}t\exp(q_{ii}t)\,,$$

where the final inequality follows by simply considering the case of no jumps. We can then apply the bounds from Assumption 6.1.5 to complete the inequality.

6.5.2 Proof of Theorem 6.1.12

We recall from [Wil67], [TC03] that for a square matrix M whose elements depend on a vector of parameters $\{\lambda_1, \dots \lambda_r\}$ (for $r \in \mathbb{N}$), the following identity holds

$$\frac{\partial e^{M(\lambda)t}}{\partial \lambda_i} = \int_0^t e^{(t-u)M(\lambda)} \frac{\partial M(\lambda)}{\partial \lambda_i} e^{uM(\lambda)} \mathrm{d}u\,, \tag{6.5.1}$$

for all $i \in \{1, \dots, r\}$. Let $\mu, \nu, \alpha, \beta \in \{1, \dots, h\}$. Recalling Proposition 6.1.2, differentiating $\mathbb{E}_Q[K_{\mu\nu}(t)|y]$ w.r.t. $q_{\alpha\beta}$ yields,

$$\begin{aligned}\frac{\partial}{\partial q_{\alpha\beta}}\mathbb{E}_Q[K_{\mu\nu}(t)|y] = &\sum_{s=1}^{n-1} -(e^{Q(t_{s+1}-t_s)})^{-2}_{y_s,y_{s+1}} \left(\frac{\partial}{\partial q_{\alpha\beta}} e^{Q(t_{s+1}-t_s)}\right)_{y_s,y_{s+1}} (e^{C_\gamma^{(\mu\nu)}(t_{s+1}-t_s)})_{y_s,h+y_{s+1}} \\ &+ (e^{Q(t_{s+1}-t_s)})^{-1}_{y_s,y_{s+1}} \left(\frac{\partial}{\partial q_{\alpha\beta}} e^{C_\gamma^{(\mu\nu)}(t_{s+1}-t_s)}\right)_{y_s,h+y_{s+1}} .\end{aligned}$$

Note that although the expected value of K only depends on individual elements of the matrix and not the full matrix, we are still able to use the differentiation result since $A_{ij} = e_i^\intercal A e_j$. Hence, from (6.5.1) we obtain,

$$\begin{aligned}\frac{\partial}{\partial q_{\alpha\beta}}\mathbb{E}_Q[K_{\mu\nu}(t)|y] = &\sum_{s=1}^{n-1} -(e^{Q(t_{s+1}-t_s)})^{-2}_{y_s,y_{s+1}} \left(\int_0^t e^{(t-u)Q} \frac{\partial Q}{\partial q_{\alpha\beta}} e^{uQ} \mathrm{d}u\right)_{y_s,y_{s+1}} (e^{C_\gamma^{(\mu\nu)}(t_{s+1}-t_s)})_{y_s,h+y_{s+1}} \\ &+ (e^{Q(t_{s+1}-t_s)})^{-1}_{y_s,y_{s+1}} \left(\int_0^t e^{(t-u)C_\gamma^{(\mu\nu)}} \frac{\partial C_\gamma^{(\mu\nu)}}{\partial q_{\alpha\beta}} e^{uC_\gamma^{(\mu\nu)}} \mathrm{d}u\right)_{y_s,h+y_{s+1}} .\end{aligned}$$

Clearly, since $q_{\alpha\beta}$ appears twice in Q,

$$\frac{\partial Q}{\partial q_{\alpha\beta}} = e_\alpha e_\beta^\intercal - e_\alpha e_\alpha^\intercal, \qquad \text{and} \qquad \frac{\partial C_\gamma^{(\mu\nu)}}{\partial q_{\alpha\beta}} = \begin{bmatrix} e_\alpha e_\beta^\intercal - e_\alpha e_\alpha^\intercal & e_\mu e_\nu^\intercal \delta_{\mu\alpha}\delta_{\nu\beta} \\ 0 & e_\alpha e_\beta^\intercal - e_\alpha e_\alpha^\intercal \end{bmatrix} .$$

Then, by [VL78] (and the proof of Proposition 6.1.2), we can solve these integrals explicitly to obtain,

$$\begin{aligned}\frac{\partial}{\partial q_{\alpha\beta}}\mathbb{E}_Q[K_{\mu\nu}(t)|y] = &\sum_{s=1}^{n-1} -(e^{Q(t_{s+1}-t_s)})^{-2}_{y_s,y_{s+1}} \left(e^{C_\eta^{(\alpha\beta)}(t_{s+1}-t_s)}\right)_{y_s,h+y_{s+1}} (e^{C_\gamma^{(\mu\nu)}(t_{s+1}-t_s)})_{y_s,h+y_{s+1}} \\ &+ (e^{Q(t_{s+1}-t_s)})^{-1}_{y_s,y_{s+1}} \left(e^{C_\psi^{(\alpha\beta,\mu\nu)}(t_{s+1}-t_s)}\right)_{y_s,3h+y_{s+1}} ,\end{aligned}$$

again $C_\eta^{(\alpha\beta)}$ and $C_\psi^{(\alpha\beta,\mu\nu)}$ are as defined in the Theorem's statement.

Therefore, we have a closed form expression for the derivative of expected jumps w.r.t. $q_{\alpha\beta}$. Applying a similar argument for the expected holding time we obtain,

$$\begin{aligned}\frac{\partial}{\partial q_{\alpha\beta}}\mathbb{E}_Q[S_\mu(t)|y] = &\sum_{s=1}^{n-1} -(e^{Q(t_{s+1}-t_s)})^{-2}_{y_s,y_{s+1}} \left(e^{C_\eta^{(\alpha\beta)}(t_{s+1}-t_s)}\right)_{y_s,h+y_{s+1}} (e^{C_\phi^{(\mu)}(t_{s+1}-t_s)})_{y_s,h+y_{s+1}} \\ &+ (e^{Q(t_{s+1}-t_s)})^{-1}_{y_s,y_{s+1}} \left(e^{C_\omega^{(\alpha\beta,\mu)}(t_{s+1}-t_s)}\right)_{y_s,3h+y_{s+1}} ,\end{aligned}$$

where $C_\omega^{(\alpha\beta,\mu)}$ is as defined in the Theorem. Combining these yields the required result.

6.5.3 Proof of Theorem 6.1.14

The proof relies on the multivariate delta method, see [LC98, Theorem 8.16] for example.

Proposition 6.5.1 (Delta Method). *Let $(X_{1\nu}, \dots, X_{s\nu})$, $\nu = 1, \dots, n$, be n independent s-tuples of random variables with $\mathbb{E}[X_{i\nu}] = \xi_i$ and $\operatorname{cov}(X_{i\nu}, X_{j\nu}) = \sigma_{ij}$. Let $\bar{X}_i$ denote the empirical mean, $\bar{X}_i := \sum_\nu X_{i\nu}/n$, and suppose that h is a real-valued function of s arguments with continuous first partial derivatives. Then,*

$$\sqrt{n}\Big(h(\bar{X}_1, \dots, \bar{X}_s) - h(\xi_1, \dots, \xi_s)\Big) \xrightarrow{Dist} \mathcal{N}(0, v^2), \quad v^2 = \sum_i \sum_j \sigma_{ij} \frac{\partial h}{\partial \xi_i}\frac{\partial h}{\partial \xi_j}, \quad \textit{provided } v^2 > 0.$$

We now have the necessary result to give the proof.

Proof of Theorem 6.1.14. The assumption of asymptotic normality implies the expectation and covariance assumption of Proposition 6.5.1. Moreover, it follows from standard results in likelihood based inference that $\sigma \approx -H(\hat{Q})^{-1}$ (see [Kni00, Chapter 5.4]).

For the partial derivatives of the probability matrix, it follows immediately by arguments in Section 6.1.2. Also note that this representation implies the first partial derivatives of p_{ij} exist and are continuous.

To complete the proof we must show the RHS of (6.1.19) is strictly positive. Firstly, at a maximum H is negative definite (hence H^{-1} is also negative definite), therefore it is enough to have that $\partial p_{ij}/\partial V_{\hat{Q}} \neq 0$ around the MLE. Observing the latter is one of the theorem's assumptions concludes the proof. □

Chapter 7

Non-Markov Setting

The aim of this chapter is to generalise away from the Markov assumption (in a sensible way). We recall that the Markov assumption arises naturally when using the TPM data in Chapter 6, hence here we consider the full dataset, which we describe in Section 7.1.

Our Contribution. In the setting of continuously observed data, we propose a tractable and parsimonious model that captures the non-Markovian phenomenon of *rating momentum*. We provide a calibration procedure and several comparative tests based on Moody's corporate credit ratings data set (see Section 7.1). Most notable is the difference between empirical, Markov (CTMC) and non-Markov (our model) estimates of probabilities of default: we observe in several cases the Markov model under or overestimates the probabilities of default empirically observed, while the non-Markov model provides better agreement.

This chapter is organised as follows. In Section 7.1 we overview the data paradigms and describe the data we work with. In Section 7.2 we use Moody's corporate credit ratings data set to test for non-Markovianity and calibrate the proposed non-Markov model; we also give due attention and discuss the effect of adding momentum in the estimation of default probabilities.

7.1 Data description

To illustrate the statistical methods we develop in this work, we use the proprietary *Moody's corporate credit ratings data set*, which comprises continuous-time observations for 17097 entities (companies) in the time from Jan 1, 1987 to Dec 31, 2017. Through the remainder of the article we refer to this as the "*Moody's data set*". Some of the discrete data is available publicly but the full data set is proprietary and must be purchased. Other papers such as [CHL04] also use the full Moody's data set.

The rating categories Moody's data set are depicted in decreasing order of rating quality as "Aaa", "Aa1", "Aa2", "Aa3", "A1", "A2", "A3", "Baa1", "Baa2", "Baa3", "Ba1", "Ba2", "Ba3", "B1", "B2", "B3", "Caa1", "Caa2", "Caa3", "Ca", "C". We define "C" as the default category. The refinements "1", "2" and "3" shall be referred to as *modifiers* in the following. The ratings "Aaa" to "Baa3" are the so-called "Investment Grade" block while the ratings "Ba1" to "Ca" form the "Speculative Grade" block.

A standard data aggregation arrangement is to aggregated all modifiers within their rating class. For instance, we group "Aa1", "Aa2", "Aa3" as "Aa" and so on to obtain the following categories in decreasing credit quality: "Aaa", "Aa", "A", "Baa", "Ba", "B", "Caa", "Ca" and "C" (Default Category). We shall use the standard aggregation unless otherwise stated.

As described there are two data paradigms, a discrete (missing) and continuous (full). In Section 6.4 we constructed annually discretised rating transition matrices from this data, and one is led to use a Markov model. The remainder of this chapter focuses on the full dataset and its richness allows one to expand the scope to non-Markov models.

7.2 Extending Markov Processes to Capture Rating Momentum

In Chapter 6 we highlighted many good features of the EM algorithm, namely, one can derive closed form expressions for the errors. However, the EM algorithm does not generalise well. One quickly runs into difficulty when using models that have more complex likelihoods. This is indeed the case when we generalise to point processes.

Before detailing the model let us start by showing that the data contains non-Markov features.

7.2.1 Testing for non-Markovian phenomena

The problem of testing whether a time series satisfies the Markov property is a well known problem. A robust test is presented in [CH12] where the Markov property is tested against a general non-parametric hypothesis. Although this approach is general, the result often only informs us whether the Markov assumption holds or not i.e. we do not typically learn the specific nature of the non-Markovianity. As we look to test specifically for the effect of so-called rating momentum we apply a similar test to that described in [LS02].

In [LS02]'s analysis of Standard and Poor's rating data set, the authors tested and showed the presence of rating momentum. For consistency and completeness we show that rating momentum is also present in Moody's data set. The test follows a standard semi-parametric hazard model approach developed in [AHK91] (see also [ABGK12]). The basic idea is to test whether the intensity (from leaving the state) is influenced by previous transitions, that is, we model the intensity for any given firm, n in state i as,

$$\lambda_{in}(t) = q_i(t)\exp(cZ_n(t)),$$

where q is an unspecified "baseline" intensity[*], Z contains information relating to the firm and c is the coefficient we estimate. One important point here is that we are often dealing with censored observations (many firms stop being rated after a while), hence using hazard models is useful since we have access to the theory of partial likelihoods which can handle censored observations, see [CO84]. One can then for example set the covariate Z as,

$$Z_n(t) = \begin{cases} 1, & \text{if firm } n \text{ was downgraded to its current state,} \\ 0, & \text{otherwise.} \end{cases}$$

Hence in this setting the Markov assumption is equivalent to the null hypothesis $c = 0$. The general statistical framework including fitting c by maximising the partial likelihood is covered in [AHK91] and [LS02, Appendix A], but we do not discuss these further here.

In [LS02] the authors test many different phenomena across a whole host of ratings, by varying the Z above[†]. However, as it is commonly accepted that rating momentum exists in the data here we only look to show a basic form of rating momentum. Essentially if the previous move was an downgrade/upgrade is there a statistically significant change in the time spent in the current state (signifying the process is not Markov).

The result can be seen in Table 7.1 – we can see a statistically significant downward momentum effect but no significant upward momentum behaviour in the Moody's data. Recall that a positive coefficient implies that the intensity increases i.e. in the case of a downgrade the firm stays in the next state for less time. These findings are consistent with those of [LS02].

	coefficient	p-value
downward momentum	0.33010	<0.0001
upward momentum	-0.01487	0.68153

Table 7.1: Likelihood ratio test for downward and upward momentum.

[*] Observe that we are not assuming that the baseline is time homogeneous in the test.

[†] One advantage of using this set up is that very little changes when testing for different phenomena. Namely all formulas stay the same the only changes are the Z and refitting c.

7.2.2 Our Proposed Rating Momentum Model

As one can see from Table 7.1, there is very strong evidence that downward momentum exists in the data. Let us now describe a methodology using *marked point processes* (see Section 5.4) that can capture this effect.

To incorporate rating momentum into such models we take inspiration from Hawkes processes and change the intensity of the model. The basic idea is to start with a CTMC (with generator matrix Q), which acts as a baseline intensity, then to that add a non-Markov component which is a self-excitation intensity that decays exponentially. That is, any downgrade observed increases the intensity of then future downgrades for a certain while. We also introduce two types of momentum, one if the company downgrades from investment grade (Baa and better) and one if from a speculative grade (this modelling choice is further discussed in Section 7.2.4 and 7.2.5). Using the same notation as before, given a state space $\{1, \dots, h\}$ such that state h (default) is absorbing. This will constitute our mark space and we model the intensity of the stochastic process X at time t as follows,

$$\lambda_g(t) = \sum_{j=1}^{h-1} q_j \mathbb{1}_{\{X(t)=j\}} + \sum_{m=1}^{2} \sum_{\tau \in \tau_m(t)} \beta_m \alpha_m e^{-\beta_m (t-\tau)} ,$$

where m denotes investment or speculative downgrade and $\tau_m(t)$ is the set of jump times (of type m) that influence the momentum at time t and α_m and β_m correspond to the intensity and memory of the "momentum" in each case. In this set up we add only four parameters to the $\approx (h-1)^2$ parameters of the CTMC case; the effectiveness of this parsimony is substantiated below. To the best of our knowledge, no other model we are aware of captures the momentum effect so simply. Further parameters and extensions can be introduced, nonetheless, we focus on this model and its analysis is found in Section 7.2.3 and 7.2.5.

The following modelling assumptions (although most of these can be easily lifted and extended to further settings), we believe these are reasonable and keep the model parsimonious.

1. We only consider downward momentum, there is no momentum for upward movement. Since upward momentum is not as statistically significant.

2. There are two types of momentum, an investment and speculative. Companies being downgraded from the investment grades (numerically these are the ratings from 1 to $(h-1)/2$) feel the investment momentum and remaining downgrades the speculative momentum.

3. Finally (not easy to remove) no points occurred prior to time 0, the so-called edge effects. This essentially says that companies do not have momentum when they are initially rated.

Remark 7.2.1 (Prudent Estimation)**.** *Since we only consider momentum as a purely negative effect, if we assume a company has no momentum when it initially does, this will give us more conservative numbers for downgrades. Therefore in calibration, if one does not use a full history of a company's rating change, the model will be more prudent.*

With these assumptions let us define the mark distribution. We take the following marked distribution (for $X(t_i) \in \{1, \dots, h-1\}$, t_i is the time of the ith jump),

$$f(X(t_i^-)|t_i) = \begin{cases} \dfrac{\sum_{j,k=1}^{h} q_{jk} \mathbb{1}_{\{X(t_i^-)=j,\ X(t_i)=k\}} + \frac{1}{N_j} \sum_{m=1}^{2} \sum_{\tau \in \tau_m(t_i)} \beta_m \alpha_m e^{-\beta_m (t_i - \tau)}}{\lambda_g(t_i)} & \text{(downgrade)}, \\ \dfrac{\sum_{j,k=1}^{h} q_{jk} \mathbb{1}_{\{X(t_i^-)=j,\ X(t_i)=k\}}}{\lambda_g(t_i)} & \text{(upgrade)} , \end{cases}$$

where we denote by t_i^- the time immediately prior to the ith jump and N_j is the number of states one can downgrade to i.e. $N_j = \sum_{k>j} \mathbb{1}_{\{q_{jk}>0\}}$. Substituting the intensity and mark distribution into (5.4.2), gives the likelihood as,

$$
\begin{aligned}
L = \prod_{i=1}^{N_g(T)} \Bigg(\Big(\sum_{j,k=1}^{h} q_{jk} \mathbb{1}_{\{X(t_i^-)=j,\ X(t_i)=k\}} + \frac{1}{N_j} \sum_{m=1}^{2} \sum_{\tau \in \tau_m(t_i)} \beta_m \alpha_m e^{-\beta_m (t_i - \tau)} \Big) \mathbb{1}_{\{X(t_i) > X(t_i^-)\}} \\
+ \sum_{j,k=1}^{h} q_{jk} \mathbb{1}_{\{X(t_i^-)=j,\ X(t_i)=k\}} \mathbb{1}_{\{X(t_i) < X(t_i^-)\}} \Bigg) \\
\times \exp\Big(- \int_0^T \sum_{j=1}^{h-1} q_j \mathbb{1}_{\{X(u)=j\}} + \sum_{m=1}^{2} \sum_{\tau \in \tau_m(u)} \beta_m \alpha_m e^{-\beta_m (u-\tau)} \mathrm{d}u \Big) . \qquad (7.2.1)
\end{aligned}
$$

Note that the likelihood is for the information from one company. We can construct the likelihood of multiple companies by taking the product but it is worthwhile noting that this assumes independence among companies. This is unlikely to be true due to business cycles etc, however, these correlated systemic effects can be introduced into risk modelling using methods from [MW07]. Hence we concentrated purely on the idiosyncratic effect of rating momentum.

The integral involving the momentum (last integral in (7.2.1)) can be simplified, to

$$
\int_0^T \sum_{\tau \in \tau_m(u)} \beta_m \alpha_m e^{-\beta_m (u-\tau)} \mathrm{d}u = \sum_{\tau \in \tau_m(T)} \alpha_m \Big(1 - e^{-\beta_m (T-\tau)} \Big) .
$$

This likelihood is complex and there appears to be no real simplification as is done in the CTMC case, the main reason for this is the time and history dependence amongst jumps for which handy relations of the form $q_{ij}^{K_{ij}}$ are no longer possible. We proceed forward by relying on Markov Chain Monte Carlo (MCMC) techniques to estimate the parameters.

7.2.3 An MCMC calibration algorithm for the model

In the TPM (using CTMC) setting considered in [BS05], [BS09] and [dRS17] the data augmentation step for CTMC was costly making the algorithm extremely slow compared to the other algorithms. In our setting, we have access to a complete data set and this expensive step is avoided. Moreover, the likelihood we deal with is complex and thus MCMC is one of the few methods that can deliver reasonable estimations.

The basic set up of MCMC is to estimate parameter(s) θ through its posterior distribution given some data D, typically denoted $\pi(\theta|D)$. In general, one cannot access this posterior distribution and direct Monte Carlo simulation is not possible as one does not know the normalising constant. MCMC gets around this by observing through Bayes' formula that,

$$
\pi(\theta|D) \propto L(D;\theta)\pi(\theta) ,
$$

where L is the likelihood and $\pi(\theta)$ is the prior distribution of θ. It is then possible to sample from this distribution using the Metropolis-Hastings algorithm with some proposal distribution.

Let X denote the set of all company transitions, we are interested in obtaining the joint distribution, $\pi(Q, \alpha, \beta|X)$ where Q is the matrix with the baseline intensities and jump probabilities (has the same form as a generator of a CTMC) and $\alpha := (\alpha_1, \alpha_2)$, $\beta := (\beta_1, \beta_2)$ are the momentum parameters. Since we assume Q, α and β to be independent, Bayes' theorem implies that,

$$
\pi(Q, \alpha, \beta|X) \propto \pi(X|Q, \alpha, \beta)\pi(Q)\pi(\alpha)\pi(\beta) = L\pi(Q)\pi(\alpha)\pi(\beta) ,
$$

where L is the likelihood defined in (7.2.1). The full conditional distribution of each parameter is obtained by conditioning on knowledge of all other parameters.

For the priors, firstly for Q, we assume that the initial transitions carry no momentum hence we can set the prior as the CTMC maximum likelihood estimate (MLE) based on the initial transitions. We therefore set the prior as exponential with the mean being the MLE. For α and β, we use a Gamma random variable with a reasonable variance as the prior. The intuition is that we have far less knowledge for these parameters but do not expect them to be zero or too large.

The next issue we tackle is the simulation from the full conditional distribution. Dealing with the parameters of the model first, their full conditional distributions are clearly not standard distributions so we use the single-component Metropolis Hastings algorithm. As always with Metropolis Hastings we need to define a good proposal function. In order to avoid a high number of rejections we take our proposal as a Gamma random variable with mean as the current step and a small variance. This in effect creates a random walk type sampling scheme that is always nonnegative. Therefore if we denote the set of parameters by γ and the proposal distribution by ψ (which can depend on the current parameters), the nth step acceptance probability of a proposed point γ_s given the current γ'_s is given by,

$$\frac{\pi(X|\gamma_s, \gamma_{n,-s})\pi(\gamma_s)\psi(\gamma'_s|\gamma_s)}{\pi(X|\gamma'_s, \gamma_{n,-s})\pi(\gamma'_s)\psi(\gamma_s|\gamma'_s)},$$

where $\gamma_{n,-s}$ denotes the set of parameters at the nth update not including the s parameter.

Model Calibration

Now that we have the necessary tools, we can calibrate our model using Moody's data set. Running 11000 MCMC iterations (taking 1000 burn in) we obtain the following results[‡].

$$Q = \begin{pmatrix} -0.0869 & 0.0836 & 0.0031 & 0 & 0.0002 & 0 & 0 & 0 & 0 \\ 0.0117 & -0.1088 & 0.0942 & 0.0025 & 0.0003 & 0.0001 & 0 & 0 & 0 \\ 0.0006 & 0.0240 & -0.0938 & 0.0666 & 0.0017 & 0.0007 & 0.0002 & 0 & 0 \\ 0.0002 & 0.0016 & 0.0387 & -0.0947 & 0.0496 & 0.0040 & 0.0006 & 0.0000 & 0 \\ 0.0001 & 0.0006 & 0.0033 & 0.0636 & -0.1774 & 0.1060 & 0.0037 & 0.0001 & 0 \\ 0.0000 & 0.0003 & 0.0012 & 0.0035 & 0.0503 & -0.1610 & 0.1012 & 0.0040 & 0.0004 \\ 0 & 0.0002 & 0.0001 & 0.0013 & 0.0048 & 0.1028 & -0.1976 & 0.0622 & 0.0261 \\ 0 & 0 & 0.0018 & 0.0029 & 0.0050 & 0.0447 & 0.1346 & -0.2838 & 0.0948 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

and for the momentum parameters,

$$\alpha = (0.031, 0.1291) \quad \text{and} \quad \beta = (3.5234, 1.7095)$$

One interesting observation arising from calibration is the difference in momentum parameters for the investment and speculative downgrades. There is apparently more momentum in the speculative downgrades than in the investment downgrades, namely, the momentum intensity is larger and lasts longer in speculative grades. One possible explanation for this is that investment grade companies are downgraded far more readily as they are monitored more closely. Therefore such companies being downgraded does not imply as much "turmoil" as it does for non investment companies. Another explanation may be due to economic effects that influence investment and non investment grade companies, for example, investment grade companies are influenced more by systemic factors than idiosyncratic factors [Cou08, pg 175], and of course our momentum model is purely idiosyncratic.

7.2.4 Bayesian Information Criterion

Let us give some justification for the use of this model. We have argued that a point process style model is a strong choice and in an effort to keep the model as robust and simple as possible we added four extra "momentum parameters" (with relation to the CTMC model). We believe four to be the optimal choice due to the fact that only adding two parameters does not yield as good a fit to what we observe and adding parameters to every rating does not seem appropriate since we do not have enough transitions across all ratings to obtain a good fit.

As we have access to the full data, one can also simply calculate the MLE Q matrix in the Markov setting. Therefore we can test our momentum model against the purely Markov model.

The Markov model is a particular case of our momentum model, hence a priori the non-Markov model stands to fit the data better. The question is if one is actually capturing the data

[‡] The MCMC algorithm, written in MATLAB, took ≈ 8.5 hours to run on a Intel Xeon E7-4660 v4 2.2GHz processor.

better or over fitting. To do this we calculate the Bayesian Information Criterion (BIC) which is a common test used in statistics for model selection and is known to penalise model complexity more than other statistical tests, such as Akaike information criterion (see [CH08, Chapter 3]). We believe this makes BIC a good test to justify our more complex model. The BIC for a model M can be written as (some authors use the negative of this)

$$\text{BIC}(M) = 2\log\big(L(M|D)\big) - \log(n)\text{dim}(M)\,,$$

where n refers to the number of data points and $\dim(M)$ is the number of parameters in the model. From a given set of models, the model with the largest BIC is taken as the best. Naturally the indicator of how much "better" one model is over another is the difference in the BIC, where a BIC difference strictly greater than 10 is taken as very strong evidence of the model superiority.

	BIC
Difference	$138.5 \gg 10$

Table 7.2: The BIC difference between the non-Markov and Markov model on the Moody's dataset.

The result in Table 7.2 provides us with confidence that our non-Markov model captures reality better without over fitting and with sufficient parsimony with relation to the Markov (CTMC) one.

7.2.5 Examples and testing

Probabilities of default as maps of time: Markov Vs. non-Markov. One important aspect of the non-Markov theory is how it impacts the TPM and transition probabilities one estimates. In the standard Markov set up the TPM is calculated using Theorem 5.1.3. In the non-Markov set up we do not have such a simple relation, hence we are forced to use Monte Carlo techniques, i.e. simulate multiple realisations according to our model and estimate the corresponding probabilities (we used 10^8 companies in each rating).

It is of particular interest to understand how the probabilities of default are altered by this model change. Using the calibrated model, Figure 7.1 details the probabilities of defaults for the various ratings as maps of time.

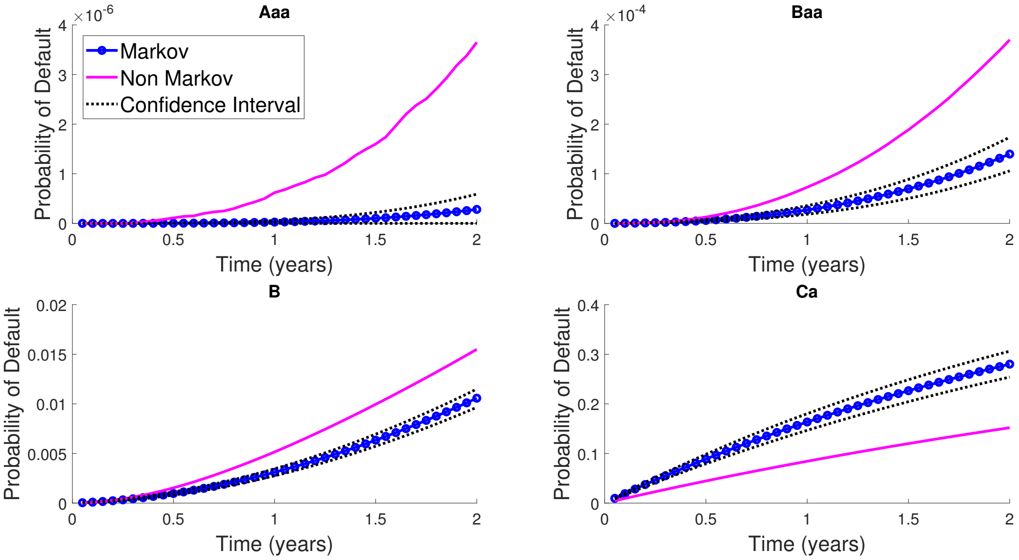


Figure 7.1: The probability of default given by each model for various ratings as a function of time. For the Markov model we have also attached the corresponding 95% confidence intervals.

The first observation one can make is that even including the confidence intervals[§] for the Markov model, the non-Markov model produces higher probabilities of default, except for the lowest ratings (the non-Markov default probability is also lower for rating Caa). The reason for this is precisely the non-Markovianity in the data. In a Markov framework all companies in the same rating are treated the same, consequently, it is unlikely that an investment grade company will continue to downgrade quickly while the non-Markov model allows for this.

On the other hand, many companies enter rating Ca before defaulting, hence in the momentum model many companies in this rating are carrying an extra term making default more likely. This implies we can account for a larger number of defaults and keep the Q matrix stable. This is not the case in the Markov model and thus to produce enough defaults one makes the Q matrix less stable.

Remark 7.2.2. *Although MCMC allows one to easily construct credible intervals for the model parameters there does not appear to be a computationally feasible way to transfer the errors in our non-Markov model parameters to the probabilities of default. We will discuss this point further in the conclusion.*

Probabilities of default: Empirical Vs. Markov Vs. non-Markov. To test how reliable these results are we can compare one year probabilities of default as estimated from each calibrated model compared to that we actually observe from the data. In order to do this we fix some time horizon, T (one year here) and consider all companies that have either defaulted or not withdrawn by the time horizon. We then build an empirical TPM over this horizon based on the company's rating at time zero and T. Concentrating solely on probabilities of default we obtain the results in Table 7.3.

Ratings	Investment Grade				Speculative Grade			
Model	Aaa	Aa	A	Baa	Ba	B	Caa	Ca
Empirical	0	0	0	0.0004000	0.0005	0.0012	0.0064	0.0563
non-Markov	1×10^{-6}	4×10^{-6}	0.0000125	0.0000734	0.0011	0.0052	0.0298	0.0845
Markov	3×10^{-8}	2.5×10^{-7}	4.86×10^{-7}	0.0000271	0.0002	0.0031	0.0407	0.1635

Table 7.3: Comparing one year probability of defaults of each model against the empirical observations.

The results in the table are interesting because the change in model makes some default probabilities higher and others lower. Starting with the investment grade, unfortunately we do not have enough data to fully assess default probabilities at this level. The only grade a default within a year is observed is Baa, it is higher than both models predict. One reason the momentum model may not capture this probability as well is the way we have set up the momentum parameters i.e. an investment and speculative set, and Baa is at the turning point. Comparing the Markov and non-Markov, it is of course unsurprising that our model makes investment grade defaults more likely.

For the speculative grades, one observes that Ca and Caa firms have lower one year default probability in the non-Markov model and these estimates are closer to the empirical observations. This is exactly due to the reason mentioned previously, companies downgrading into Ca and Caa "poison" the data in the Markov setting. Implying that in a Markov world a company initially rated Caa or Ca is viewed to be more risky than it actually is.

The difference between the models may have a large impact on a bank's capital requirements for regulation. Although the non-Markov model makes most ratings more risky than the Markov model, we feel it provides a more accurate reflection of default risk.

Remark 7.2.3 (Limitations from censored data)**.** *Unfortunately we are limited to small time horizons here due to censored data. Namely, since default is absorbing as soon as a company defaults we keep that information up to the terminal time. However, many companies are only rated over a few years and therefore if we look at empirical TPMs over longer horizons they are built with less (non default) data. Since we do not want to use the Markov assumption there does not appear to be*

[§] In this case we have calibrated the Markov model using the full data. Hence we have no expected values appearing in the MLE estimate and the Hessian is simply a diagonal matrix. One can then again use the Delta method as before but with this simpler Hessian.

a way to incorporate this lost data. Therefore we can only obtain "accurate" numbers on short time scales.

7.3 Conclusion and Outlook

In Chapter 6 we obtained closed form expressions for the expected number of jumps and holding times of a CTMC with an absorbing state, given discrete observations. We then used these representations to

derive a closed form expression for the Hessian of the likelihood which is crucial in calculating Wald confidence intervals. Although errors at the level of the generators are useful, from a practitioner standpoint errors at the level of transition probabilities are more important and through the delta we provide a relatively simple formula to obtain these. This coupled with stronger convergence has elevated the EM algorithm to be the optimal algorithm to tackle this particular problem.

Across the battery of tests carried out, the EM algorithm outperforms competing algorithms. The EM is a tractable algorithm, slower than the deterministic algorithms but still several orders of magnitude faster than the Markov-Chain Monte-Carlo alternatives (Table 6.2). The statistical algorithms (EM and MCMC) embed a strong robustness property for the estimator contrary to the deterministic algorithms, i.e. the likelihood is far less sensitive to small changes in the underlying TPM. On the more practical side, Figure 6.5 highlights that for lower ratings algorithms produce essentially the same estimates for the probabilities of default while a palpable difference emerges at higher ratings.

In a data paradigm, in Chapter 7 we have shown the significance of being able to capture non-Markov effects in rating transitions. Comparing against empirical probabilities of default, one finds a tendency for the Markov chain model to overestimate on some speculative grades and underestimate on investment grades. We address this issue by providing a parsimonious model that better captures default probabilities, where empirically observed. Moreover, the non-Markov model points toward significantly higher probabilities of default in the investment grades, where such values are not empirically observed, thus making it more prudent. It is our belief that the model we present provides a more accurate view of reality and hence should be considered in credit risk modelling. These observations further highlight the importance of understanding so-called *model risk* and its potential impact in risk calculations.

One issue highlighted in this chapter was the expensiveness of calculating errors (credible intervals) in the non-Markov model. As discussed obtaining the intervals at the level of the model parameters is straightforward[¶], however, transferring this to the transition probabilities requires us to numerically calculate (via Monte Carlo) a transition matrix for each realised set of model parameters. In essence the problem is a Monte Carlo inside a Monte Carlo which is often computationally expensive. A future line of research would therefore be to develop a more efficient method to go from the parameter intervals to transition probabilities, either by optimally selecting values or via importance sampling (this could work especially well when one is only interested in default probabilities).

[¶] In the MCMC set up we already have the set of iterations, one therefore simply calculates the smallest interval that contains x% (say 95) of these realisations.

Part III

Numerical Algorithms for McKean-Vlasov SDEs

Chapter 8

Preliminaries

This part is based on the author's joint work with dos Reis and Tankov [RST18] and dos Reis and Engelhardt [RES18].

We recall that this chapter involves the study of McKean Vlasov Stochastic Differential Equations (MV-SDEs). As remarked in the introduction, compared to standard SDEs, there are far less numerical techniques available for MV-SDEs. This is a major issue due to the growing popularity of these equations and the added computational complexity involved in their simulation.

Motivated by the work on standard SDEs we address the challenge of deriving numerical algorithms for MV-SDEs with superlinear growing drifts. Chapter 9 focuses on the convergence of explicit and implicit numerical schemes for such MV-SDEs, while Chapter 10 focuses on variance reduction, in particular we propose two importance sampling algorithms. Importance sampling is a variance reduction technique whereby one changes the measure under which simulations are carried out, this is far more involved in the case of MV-SDEs since the coefficients themselves depend on this measure.

Let us review the core concepts required for this work and present our results in Chapters 9 and 10.

8.1 Wasserstein Metric

As MV-SDEs depend on measures (via their own law) in order to do calculations involving measures we must introduce a notion of distance in this space. The standard choice to determine distances between measures in measure spaces (which is useful for the MV-SDEs framework) is the so-called Wasserstein distance, see [Vil08, Chapter 6] for details.

Given the measurable space $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, we denote by $\mathcal{P}(\mathbb{R}^d)$ the set of probability measures on this space, and write $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ if $\mu \in \mathcal{P}(\mathbb{R}^d)$ and, $\int_{\mathbb{R}^d} |y|^2 \mu(\mathrm{d}y) < \infty$. We then have the following metric on the space $\mathcal{P}_2(\mathbb{R}^d)$, the so-called Wasserstein metric (or distance) for $\mu,\ \nu \in \mathcal{P}_2(\mathbb{R}^d)$,

$$W^{(2)}(\mu,\nu) = \inf_{\pi} \Big\{ \Big(\int_{\mathbb{R}^d \times \mathbb{R}^d} |x-y|^2 \pi(\mathrm{d}x, \mathrm{d}y) \Big)^{1/2} \ : \ \pi \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d) \text{ with marginals } \mu \text{ and } \nu \Big\} . \tag{8.1.1}$$

There is of course no need to take $p = 2$ and have the measures defined on $\mathbb{R}^d$, one can generalise this to any $p \in [1, \infty)$ and consider a Polish metric space $(\mathcal{X}, d)$, see [Vil08, Definition 6.1]. However, this is far more general than we need for our setting. Proof that the Wasserstein is indeed a metric (satisfies the axioms of distance is given in [Vil08, pg 94]).

One may ask, why Wasserstein is a good choice of metric? The Wasserstein metric features heavily in optimal transport theory and [Vil08, Chapter 6] gives motivation why that is the case. For MV-SDEs the main reason is that due to the infimum it is relatively simple to bound and is in some sense analogous to the Lipschitz condition. To see this consider the distance between the integrals

$$\int_{\mathbb{R}^d} f(x)\mathrm{d}\mu(x) \quad \text{and} \quad \int_{\mathbb{R}^d} f(x)\mathrm{d}\nu(x)\,,$$

where f is a global Lipschitz function, one can see that the distance between these integrals is bounded by the Wasserstein distance (up to a constant) through the following,

$$\begin{aligned} |\int_{\mathbb{R}^d} f(x)\mathrm{d}\mu(x) - \int_{\mathbb{R}^d} f(y)\mathrm{d}\nu(y)| &= |\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} f(x)\mathrm{d}\pi(x,y) - \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} f(y)\mathrm{d}\pi(y,x)| \\ &\leq C\int_{\mathbb{R}^d}\int_{\mathbb{R}^d} |x-y|\mathrm{d}\pi(x,y)\,, \end{aligned}$$

where π has marginals μ and ν and we have used μ and ν are probability measures to obtain the first equality. Taking infimum over π and using Jensen's inequality implies the distance between the integrals is bounded by $CW^{(2)}(\mu,\nu)$.

One of the results that is useful for us is the convergence of an empirical distribution to the true in $W^{(2)}$. Namely, consider a random variable Y with law (Borel probability measure) μ, then draw N independent samples from Y, it is known that

$$\mu^N := \frac{1}{N}\sum_{i=1}^{N} \delta_{Y^i} \to \mu \quad \text{as } N \to \infty\,,$$

weakly with probability 1 (see [Dud18, Theorem 11.4.1]). But moreover, one can also consider the convergence of this in terms of N, namely, provided $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ we wish to consider $\mathbb{E}[W^{(2)}(\mu^N,\mu)^2]$, the classical convergence result can be found in [RR98, Chapter 10.2], but more recently the convergence has been improved upon, to the following (see [CD17a, Theorem 5.8]).

Theorem 8.1.1. *Let $\mu \in \mathcal{P}_q(\mathbb{R}^d)$ for some $q > 4$, then for each $d \geq 1$ there exists a constant (dependent on d, q and μ) such that for all $N \geq 2$,*

$$\mathbb{E}[W^{(2)}(\mu^N,\mu)^2] \leq C \begin{cases} N^{-1/2} & \text{if } d < 4, \\ N^{-1/2}\log(N) & \text{if } d = 4, \\ N^{-2/d} & \text{if } d > 4. \end{cases}$$

This result is useful when one wishes to consider the so-called particle approximation of a MV-SDE converging to the true solution, this is crucial for results below and will be discussed more thoroughly.

8.2 McKean-Vlasov stochastic differential equations

Let W be an l-dimensional Brownian motion and take the progressively measurable maps $b : [0,T] \times \mathbb{R}^d \times \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}^d$ and $\sigma : [0,T] \times \mathbb{R}^d \times \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}^{d\times l}$. MV-SDEs are typically written in the form,

$$\mathrm{d}X_t = b(t, X_t, \mu_t^X)\mathrm{d}t + \sigma(t, X_t, \mu_t^X)\mathrm{d}W_t, \quad X_0 \in L_0^p(\mathbb{R}^d), \tag{8.2.1}$$

where μ_t^X denotes the law of the process X at time t, i.e. $\mu_t^X = \mathbb{P}\circ X_t^{-1}$. We make the following assumption on the coefficients throughout.

Assumption 8.2.1. *Assume that σ is Lipschitz in the sense that there exists $L > 0$ such that for all $t \in [0,T]$ and all $x, x' \in \mathbb{R}^d$ and $\forall \mu, \mu' \in \mathcal{P}_2(\mathbb{R}^d)$ we have that*

$$|\sigma(t,x,\mu) - \sigma(t,x',\mu')| \leq L(|x-x'| + W^{(2)}(\mu,\mu')),$$

and let b satisfy

1. *One-sided Lipschitz in x and Lipschitz in law: there exist $L_b,\ L>0$ such that for all $t\in[0,T]$, all $x,x'\in\mathbb{R}^d$ and all $\mu,\mu'\in\mathcal{P}_2(\mathbb{R}^d)$ we have that*
$$\langle x-x', b(t,x,\mu)-b(t,x',\mu)\rangle \le L_b|x-x'|^2 \ \text{ and } \ |b(t,x,\mu)-b(t,x,\mu')|\le LW^{(2)}(\mu,\mu').$$

2. *Locally Lipschitz with polynomial growth in x: there exists $q\in\mathbb{N}$ with $q>1$ such that for all $t\in[0,T]$, $\forall\mu\in\mathcal{P}_2(\mathbb{R}^d)$ and all $x,\ x'\in\mathbb{R}^d$*
$$|b(t,x,\mu)-b(t,x',\mu)|\le L(1+|x|^q+|x'|^q)|x-x'|.$$

Using the one-sided Lipschitz drift, a particularised version of [dRST17, Theorem 3.3] provides a result for existence and uniqueness.

Theorem 8.2.2 ([dRST17])**.** *Suppose that b and σ satisfy Assumption 8.2.1 and be continuous in time. Further, assume for some $m\ge 2$, $X_0\in L_0^m(\mathbb{R}^d)$. Then there exists a unique solution for $X\in\mathbb{S}^m([0,T])$ to the MV-SDE* (8.2.1). *For some positive constant C we have*
$$\mathbb{E}\big[\sup_{t\in[0,T]}|X_t|^m\big]\le C\Big(\mathbb{E}[|X_0|^m]+\Big(\int_0^T b(t,0,\delta_0)\mathrm{d}t\Big)^m+\Big(\int_0^T \sigma(t,0,\delta_0)^2\mathrm{d}t\Big)^{m/2}\Big)e^{CT}.$$

If the law μ^X is known beforehand, then the MV-SDE reduces to a "standard" SDE with added time-dependency. Typically this is not the case and usually the MV-SDE is approximated by a weakly interacting particle system. This argument is sometime referred to as a *decoupling argument* or *decoupling technique*, since as we take the number of particles to infinity, the dependence of one particle any other vanishes.

The interacting particle system approximation

We approximate (8.2.1) (driven by the Brownian motion W), using an N-dimensional system of interacting particles. Let $i=1,\dots,N$ and consider N particles $X^{i,N}$ satisfying the SDE with $X_0^{i,N}=X_0^i$ (since the initial condition is random, but independent of other particles)
$$\mathrm{d}X_t^{i,N}=b\Big(t,X_t^{i,N},\mu_t^{X,N}\Big)\mathrm{d}t+\sigma\Big(t,X_t^{i,N},\mu_t^{X,N}\Big)\mathrm{d}W_t^i, \tag{8.2.2}$$
where $\mu_t^{X,N}(\mathrm{d}x):=\frac{1}{N}\sum_{j=1}^N\delta_{X_t^{j,N}}(\mathrm{d}x)$ and $\delta_{X_t^{j,N}}$ is the Dirac measure at point $X_t^{j,N}$, and the independent Brownian motions $W^i, i=1,\dots,N$ (also independent of the BM W appearing in (8.2.1); with a slight abuse of notation to avoid re-defining the probability space's Filtration).

Propagation of chaos. In order to show that the particle approximation is of use, one shows a pathwise propagation of chaos result. Although different types exist we are interested in strong error hence require a pathwise convergence result where we consider the system of non interacting particles
$$\mathrm{d}X_t^i=b(t,X_t^i,\mu_t^{X^i})\mathrm{d}t+\sigma(t,X_t^i,\mu_t^{X^i})\mathrm{d}W_t^i,\quad X_0^i=X_0^i\,,\quad t\in[0,T]\,, \tag{8.2.3}$$
which are of course just MV-SDEs and since the X^is are independent, then $\mu_t^{X^i}=\mu_t^X$ for all i. Under global Lipschitz conditions, one can then prove the following convergence result (see [Car16, Theorem 1.10] for example)
$$\lim_{N\to\infty}\sup_{1\le i\le N}\mathbb{E}\big[\sup_{0\le t\le T}|X_t^{i,N}-X_t^i|^2\big]=0\,.$$

All SDEs appearing below have initial condition X_0^i and we work on the interval $[0,T]$.

Remark 8.2.3 (Requirement to extend propagation of chaos)**.** *The above convergence result does not cover the non Lipschitz growth we allow for in Theorem 8.2.2. Hence we look to generalise the result, which is done in Proposition 9.1.2.*

8.3 Explicit and Implicit for methods for super linear SDEs

We have already alluded to the issue of using basic a Euler (Euler Maruyama) scheme to simulate SDEs with super linear growth. In order to make the work self contained we now look to present some of the schemes one can use in this setting for standard SDEs. Later we introduce potential MV-SDE equivalents of these schemes.

Remark 8.3.1 (Euler Scheme for non Lipschitz). *One should note that the Euler scheme is known to converge in some non globally Lipschitz settings, for example in the locally Lipschitz with linear growth,* [YM08] *show convergence. Hence we are focusing here on the case where the spatial coefficient of drift can grow faster than linearly.*

The main result proving that the Euler scheme does not work in this setting was given in [HJK11], consider a standard SDE, X with measurable coefficients b and σ such that the SDE,

$$\mathrm{d}X_t = b(X_t)\mathrm{d}t + \sigma(X_t)\mathrm{d}W_t, \quad X_0 = x_0\,,$$

has a unique strong solution for $t \in [0,T]$. For ease of presentation we shall use this SDE as a running example in this section. One can then consider an Euler approximation on this SDE by introducing the iterative sequence $Y^M_{t_k} : \Omega \to \mathbb{R}$, $k = \{0, \dots, M-1\}$, and $M \in \mathbb{N}$, where $Y^M_0 = x_0$ and,

$$Y^M_{t_{k+1}} = Y^M_{t_k} + \frac{T}{M} b(Y^M_{t_k}) + \sigma(Y^M_{t_k})\big(W_{t_{k+1}} - W_{t_k}\big).$$

This leads to the divergence result [HJK11, Theorem 1].

Theorem 8.3.2. *Assume existence of a strong solution to X and assume $\sigma(x_0) \neq 0$ and let there exist constants $C \geq 1$, $\beta > \alpha > 1$ such that for all $|x| \geq C$,*

$$\max(|b(x)|, |\sigma(x)|) \geq \frac{|x|^\beta}{C} \quad \textit{and} \quad \min(|b(x)|, |\sigma(x)|) \leq C|x|^\alpha\,.$$

Then there exists a constant $c \in (1, \infty)$ and a sequence of nonempty events $\Omega_M \in \mathcal{F}$ for all $M \in \mathbb{N}$, that satisfy $\mathbb{P}(\Omega_M) \geq c^{(-M^c)}$ and $|Y^N_T(\omega)| \geq 2^{(\alpha^{(M-1)})}$ for all $\omega \in \Omega_M$.

Moreover, if the exact solution $X \in L^p_T$ for some $p \in [1, \infty)$, then,

$$\lim_{M \to \infty} \mathbb{E}[|X_T - Y^M_T|^p] = \infty \quad \textit{and} \quad \lim_{M \to \infty} \Big|\mathbb{E}[|X_T|^p] - \mathbb{E}[|Y^M_T|^p]\Big| = \infty.$$

The point of this theorem is that although the set of $\omega \in \Omega$ where the Euler scheme attains large values is small (exponentially small), when such realisations do occur however, the values are much larger (double exponentially large). This implies that the L^1 norm is unbounded for all M. Hence we obtain not only that the Euler scheme does not converge as $M \to \infty$ but it in fact diverges. One may also note the somewhat unintuitive max and min condition on the coefficients, this essentially is to guarantee that either the drift or diffusion grows faster than linearly and the other grows slower than that.

Many SDEs exist and have unique strong solutions that satisfy the growth condition in the above theorem, see for example [Mao08, Theorem 2.3.5]. Hence in order to simulate these SDEs we are required to develop more sophisticated sampling techniques. There have been some more crude approaches to solving the problem, for example, removing paths that leave a suitably large ball as considered in [MT05]. As we shall discuss however, such approaches are not suitable for generalising away from standard the SDE setting to the MV-SDE setting due to the law dependence. We therefore concentrate on the two main approaches to tackling this problem, the explicit (tamed Euler) and implicit (backward Euler) schemes.

The original solution to the problem of simulating super linear growing SDEs was given in [HMS02], where the authors propose a so called backward Euler scheme. With our example SDE X above the backward Euler scheme (commonly referred to as the implicit scheme) is the iterative process defined by $\tilde{Y}_0^M = x_0$ and,

$$\tilde{Y}_{t_{k+1}}^M = \tilde{Y}_{t_k}^M + \frac{T}{M} b(\tilde{Y}_{t_{k+1}}^M) + \sigma(\tilde{Y}_{t_k}^M)\big(W_{t_{k+1}} - W_{t_k}\big).$$

One observes that the main difference here is that $\tilde{Y}_{t_{k+1}}^M$ appears on both sides of the equation, hence this scheme relies on solving for the fixed point. The advantage of this is one does not observe the wild oscillations that appear when using the Euler scheme in the super linear setting. This allows us to obtain strong convergence even when we have super linear coefficients. As we look to generalise such results we do not give the details here but an interested reader can consult [HMS02] and [MS13] for further details.

Although the implicit allows us to deal with SDEs that have more general coefficients the main issue with this method is the computational complexity. This is due to the appearance of $\tilde{Y}_{t_{k+1}}^M$ on both sides, which implies we are required to solve a fixed point equation at every time step. This is more expensive than standard explicit schemes and moreover scales with dimension squared (see [HJK12]). The goal is therefore to construct an explicit algorithm capable of handling a super linear (drift) coefficient. This was achieved in [HJK12] where the authors propose a so-called *tamed* Euler scheme. Since this original work other explicit schemes have been developed, see for example [Sab13] and [CJM16]. Although all schemes are based on the idea that the drift must be truncated (but the truncation is dependent on stepsize). Taking the scheme proposed in [Sab13], the tamed Euler has the form, $\bar{Y}_0^M = x_0$ and,

$$\bar{Y}_{t_{k+1}}^M = Y_{t_k}^M + \frac{T}{M}\frac{b(Y_{t_k}^M)}{1 + \left(\frac{T}{M}\right)^\alpha |b(Y_{t_k}^M)|} + \sigma(Y_{t_k}^M)\big(W_{t_{k+1}} - W_{t_k}\big),$$

where $\alpha \in (0, 1/2]$. The key feature here is that the denominator in the drift bounds the size, however, as $M \to \infty$ the bound becomes less strict. It is then possible to show this scheme converges for super linear drifts and has the advantage of being completely explicit. Again we do not give the precise convergence results here but one can consult, [HJK12], [Sab13], [CJM16] amongst others for further details and discussion of the various schemes.

8.4 Large Deviation Principles

In this section, we state and review the main results from the large deviations theory that we require. This is a small overview of the theory, for a full exposition the reader can consult texts such as [DZ10] or [DE11]. The large deviation principle (LDP) characterises the limiting behaviour, as $\epsilon \to 0$, of a family of probability measures $\{\mu_\epsilon\}$ in exponential scale on the space $(\mathcal{X}, \mathcal{B}_\mathcal{X})$, with $\mathcal{X}$ a topological space so that open and closed subsets of $\mathcal{X}$ are well-defined, and $\mathcal{B}_\mathcal{X}$ is the Borel σ-algebra on $\mathcal{X}$. The limiting behaviour is defined via a so-called rate function. We assume the probability spaces have been completed, consequently, $\mathcal{B}_\mathcal{X}$ is the complete Borel σ-algebra on $\mathcal{X}$. We have the following definition [DZ10, pg.4].

Definition 8.4.1 (Rate function)**.** *A rate function I is a lower semicontinuous mapping $I : \mathcal{X} \to [0, \infty]$ (such that for all $\alpha \in [0, \infty)$, the level set $\Psi_I(\alpha) := \{x : I(x) \le \alpha\}$ is a closed subset of $\mathcal{X}$). A good rate function is a rate function for which all the level sets $\Psi_I(\alpha)$ are compact subsets of $\mathcal{X}$. The effective domain of I, denoted D_I, is the set of points in $\mathcal{X}$ of finite rate, namely, $D_I := \{x : I(x) < \infty\}$.*

We use the standard notation: for any set Γ, $\overline{\Gamma}$ denotes the closure, Γ^o denotes the interior and finally Γ^C denotes the complement of Γ. As is standard practice in LDP theory, the infimum of a function over an empty set is interpreted as ∞. We then define what it means for this sequence of measures to have an LDP [DZ10, pg.5].

Definition 8.4.2. *A family of probability measures, $\{\mu_\epsilon\}$ with $\epsilon > 0$ satisfies the large deviation principle with a rate function I if, for all $\Gamma \in \mathcal{B}$,*

$$-\inf_{x \in \Gamma^o} I(x) \le \liminf_{\epsilon \to 0} \epsilon \log \mu_\epsilon(\Gamma) \le \limsup_{\epsilon \to 0} \epsilon \log \mu_\epsilon(\Gamma) \le -\inf_{x \in \overline{\Gamma}} I(x)\,. \tag{8.4.1}$$

It is also typical to have LDP defined in terms of a sequence of random variables Z_ϵ, in which case one replaces $\mu_\epsilon(\Gamma)$ by $\mathbb{P}[Z_\epsilon \in \Gamma]$.

The following result can be viewed as a generalisation of Laplace's approximation of integrals to the infinite dimensional setting and transfers the LDP from probabilities to expectations (see [DZ10]).

Lemma 8.4.3 (Varadhan's Lemma). *Let $\{\mu_\epsilon\}$ be a family of measures that satisfies a large deviation principle with good rate function I. Furthermore, let Z_ϵ be a family of random variables in $\mathcal{X}$ such that Z_ϵ has law μ_ϵ and let $\varphi : \mathcal{X} \to \mathbb{R}$ be any continuous function that satisfies the following integrability (moments) condition for some $\gamma > 1$,*

$$\limsup_{\epsilon\to 0} \epsilon \log \mathbb{E}\left[\exp\left(\frac{\gamma}{\epsilon}\varphi(Z_\epsilon)\right)\right] < \infty\,.$$

Then,

$$\lim_{\epsilon\to 0} \epsilon \log \mathbb{E}\left[\exp\left(\frac{1}{\epsilon}\varphi(Z_\epsilon)\right)\right] = \sup_{x\in\mathcal{X}} \left\{\varphi(x) - I(x)\right\}\,.$$

As is discussed in [GR08], one needs a slight extension to Varadhan's lemma to allow the function φ to take the value $-\infty$. The extension is proved in [GR08].

Lemma 8.4.4. *Let $\varphi : \mathcal{X} \to [-\infty, \infty)$ and assume the conditions in Lemma 8.4.3 are satisfied. Then the following bounds hold for any $\Gamma \in \mathcal{B}$*

$$\begin{aligned}\sup_{x\in\Gamma^0}\{\varphi(x) - I(x)\} &\le \liminf_{\epsilon\to 0} \epsilon \log\left(\int_{\Gamma^o} \exp\left(\frac{1}{\epsilon}\varphi(Z_\epsilon)\right) \mathrm{d}\mu_\epsilon\right)\\ &\le \limsup_{\epsilon\to 0} \epsilon \log\left(\int_{\overline{\Gamma}} \exp\left(\frac{1}{\epsilon}\varphi(Z_\epsilon)\right) \mathrm{d}\mu_\epsilon\right) \le \sup_{x\in\overline{\Gamma}}\{\varphi(x) - I(x)\}\,.\end{aligned}$$

The previous lemma allows us to control the $\liminf$ and $\limsup$ of the process even when they are not equal (as is the case in Varadhan's lemma).

8.4.1 Sample Path Large Deviation

Varadhan's lemma is useful because it turns a potentially awkward expression involving limits into a simpler optimisation problem, provided we know the corresponding rate function. For our purposes the random variable in question will be Brownian motion, hence we wish to obtain its rate function. LDP results require us to introduce some parameter ϵ (that we can take to zero), hence consider the process for a BM, W taking values in $\mathbb{R}^d$,

$$W_\epsilon(t) = \sqrt{\epsilon} W_t\,,$$

and denote by ν_ϵ the probability measure* induced by $W_\epsilon(\cdot)$. This then leads to the following result, see [DZ10, Theorem 5.2.3].

Theorem 8.4.5 (Schilder). *The family of measures $\{\nu_\epsilon\}$ in $C_0([0,T])$ satisfies a LDP with good rate function given by,*

$$I_W(\phi) = \begin{cases} \frac{1}{2}\int_0^T |\dot{\phi}(t)|^2 \mathrm{d}t, & \phi \in \mathbb{H}^d_T,\\ \infty & \textit{otherwise.}\end{cases}$$

This implies that when dealing with BM, we can now write the optimisation problem in Varadhan's lemma explicitly.

Remark 8.4.6. *There is also some interesting work connecting MV-SDEs and large deviations through there empirical measures, see* [DMG98], [BDF12] *and* [Fis14] *for example. Although related these works consider a different problem to ours hence we mention them purely for reference.*

*As one does with the Wiener measure, we can view ν_ϵ as a measure on the space of continuous functions mapping $\mathbb{R}_+$ to $\mathbb{R}^d$, which have value zero at time zero.

8.5 Importance Sampling and large deviations

Monte Carlo is an extremely useful numerical technique that is used to approximate integrals, see [Gla13] for a full discussion and its application to finance. Despite the technique being quite general it can suffer from rather poor convergence. The reason for this is the error (which is based on the variance of the estimate) only reduces with order $1/\sqrt{N}$, for N Monte Carlo samples. In view of the poor convergence of the standard Monte Carlo, it is typical to enhance the standard approach with a so-called *variance reduction* technique, *importance sampling* is one such technique and what we focus on here.

To motivate our approach we recall ideas from the pioneering works [GHS99], [GR08] and [Rob10] which establish a connection between large deviations and importance sampling. Importance sampling uses the following idea. Consider the problem of estimating $\mathbb{E}_{\mathbb{P}}[G(X)]$ where X is some random variable/process governed by probability $\mathbb{P}$. Through Radon-Nikodym theorem we can rewrite this expectation under a new measure $\mathbb{Q}$ weighted by the Radon-Nikodym derivative, thus $\mathbb{E}_{\mathbb{P}}[G(X)] = \mathbb{E}_{\mathbb{Q}}[G(X)\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}]$. Although the expectations (first moments) are the same, the variance under $\mathbb{Q}$ is,

$$\mathrm{Var}_{\mathbb{Q}}\Big[G(X)\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}\Big] = \mathbb{E}_{\mathbb{P}}\Big[G(X)^2\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}\Big] - \mathbb{E}_{\mathbb{P}}\Big[G(X)\Big]^2. \tag{8.5.1}$$

As it turns out, if one chooses $\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}} = \frac{G}{\mathbb{E}_{\mathbb{P}}[G]}$, then the variance under $\mathbb{Q}$ is zero, i.e. we have no error in our Monte Carlo simulation. Unfortunately though, in order to choose such a change of measure one would need to a priori know the value of $\mathbb{E}_{\mathbb{P}}[G(X)]$ i.e. the value we wish to estimate in the first place.

Instead one typically chooses $\mathbb{Q}$ to minimise (8.5.1) over a set of equivalent probability measures, chosen to add only a small amount of extra computation and such that the process X is easy to simulate under the new measure. Specialising to the Brownian filtration, a common choice of $\mathbb{Q}$ is the Girsanov transform, (8.5.2) where f is often taken to be a deterministic function.

For example in [TFC16] the authors develop an importance sampling procedure in the context of Gaussian random vectors through a so-called "tilting" parameter, which corresponds to shifting the mean of the Gaussian random vector via a Girsanov transform. Although this method is intuitive, it still requires estimation of the Jacobian of G w.r.t. the tilting parameter and applying Newton's method to select the optimal parameter value. These steps can be computationally expensive, and it is difficult to obtain rigorous optimality results.

Even after one has reduced the set of measures $\mathbb{Q}$ to optimise over, in general the problem of minimising (8.5.1) will not have a closed form solution. Thus we instead minimise a proxy for the variance obtained in the so-called small noise asymptotic regime as discussed in [GHS99] and [GR08]. Assuming that a Girsanov change of measure is used, we want to minimise

$$\mathbb{E}_{\mathbb{P}}\Big[G(W)^2\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}\Big] = \mathbb{E}_{\mathbb{P}}\Big[\exp\Big(2F(W) - \int_0^T f_t \mathrm{d}W_t + \frac{1}{2}\int_0^T f_t^2 \mathrm{d}t\Big)\Big], \quad \text{with } F = \log(G). \tag{8.5.2}$$

Typically G is defined as a functional of the SDE, but here with a slight abuse of notation we have redefined it as the functional of the driving Brownian motion. It is important for this type of argument that we are able to write the solution of the SDE in terms of BM as well, i.e. we can write $X_t = H(t, W_\cdot)$. Finding the optimal f by minimising (8.5.2) is in general intractable, hence an asymptotic approximation of the variance should be constructed. Let us consider,

$$\epsilon \log \mathbb{E}_{\mathbb{P}}\Big[\exp\Big(\frac{1}{\epsilon}\Big(2F(\sqrt{\epsilon}W) - \int_0^T \sqrt{\epsilon} f_t \mathrm{d}W_t + \frac{1}{2}\int_0^T f_t^2 \mathrm{d}t\Big)\Big)\Big],$$

which equals $\log$ of (8.5.2) when $\epsilon = 1$, the *small noise asymptotic approximation* is then,

$$L(f) := \limsup_{\epsilon \to 0} \epsilon \log \mathbb{E}_{\mathbb{P}}\Big[\exp\Big(\frac{1}{\epsilon}\Big(2F(\sqrt{\epsilon}W) - \int_0^T \sqrt{\epsilon} f_t \mathrm{d}W_t + \frac{1}{2}\int_0^T f_t^2 \mathrm{d}t\Big)\Big)\Big].$$

One then computes a candidate variance reduction parameter f^* by minimising $L(f)$, which can be thought of as approximating $\mathbb{E}_{\mathbb{P}}\left[G(W)^2 \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}\right]$ by $\exp(L(f))$. Crucially, L is in a form that can be evaluated using the Varadhan's lemma, i.e., we can change L into a supremum depending on the rate function. One other advantage of using this method is that it allows us to also ask whether the measure change we have selected is optimal (in some sense). The standard optimality concept in this setting is so-called *asymptotically optimal*, which we will give a precise definition of later. Moreover, the measure change we use here is a deterministic measure change arising from LDP, there is a stochastic variant of this and we shall come back to this point in Chapter 10. It is important to note, these approximations are not approximations for the original problem (calculate $\mathbb{E}_{\mathbb{P}}[G(X)]$), they are only approximations to help choose the change of measure we want to apply.

Remark 8.5.1 (Minimise Error). *This section has discussed how one can use LDP to obtain an optimisation problem to minimise variance. A different (and perhaps more natural) way to approach the problem is to use Cramér's theorem (see* [DZ10, Chapter 2.2]*) and calculate the optimal measure change in terms of optimising the corresponding rate function. This way corresponds to minimising the error between the empirical (Monte Carlo) and the true expectation.*

This is the approach adopted in [HN16] *and they show the method in simple cases. However, it is unclear how to extend this method to more general cases and therefore is not a practical alternative at present.*

8.6 Deterministic Optimal Control

Due to the fact that our variance reduction is an optimisation problem, that is, we wish to find the change of measure (the function f living in some space) which minimises our proxy for the variance. It turns out the theory we require is (deterministic) optimal control. Optimal control is essentially an extension to calculus of variation (which itself is an extension of calculus). The main difference being, optimal control allows for inequality constraints while calculus of variation requires equality constraints. These theories have had many applications and interesting results which we do not discuss here, however, one can consult [FR75], [YZ99] among others for further details.

One of the most important results from optimal control is Pontryagin's maximum principle. Roughly speaking, Pontryagin's maximum principle gives a set of differential equations that the optimal control must satisfy. Let us recall the main ideas following [YZ99, p.102]. We start with the controlled dynamical system $x(t)$ which takes the following form:

$$\begin{cases} \dot{x}(t) = b(t, x(t), u(t)), & \text{a.e. } t \in [0,T] \\ x(0) = x_0 \,, \end{cases} \tag{8.6.1}$$

where u is our "control", which is defined in a metric space (U, d) and associated to this we have a *cost functional*

$$J(u(\cdot)) = \int_0^T f(t, x(t), u(t))\mathrm{d}t + h(x(T))\,, \tag{8.6.2}$$

f is typically referred to as the *running cost* and h the *terminal cost*. We then have the following assumption.

Assumption 8.6.1. *For ease of writing we denote by $\varphi(t, x, u)$ to be any of the functions $b(t, x, u)$, $f(t, x, u)$ or $h(x)$. We then assume the following,*

- *(U, d) is a separable metric space and $T > 0$.*
- *The maps $b : [0, T] \times \mathbb{R}^n \times U \to \mathbb{R}^n$, $f : [0, T] \times \mathbb{R}^n \times U \to \mathbb{R}$ and $h : \mathbb{R}^n \to \mathbb{R}$ are measurable and there exists a constant $L > 0$ and a modulus of continuity $\eta : [0, \infty) \to [0, \infty)$ such that,*

$$\begin{cases} |\varphi(t, x, u) - \varphi(t, \hat{x}, \hat{u})| \le L|x - \hat{x}| + \eta(d(u, \hat{u})) & \forall t \in [0, T]\ x, \hat{x} \in \mathbb{R}^n,\ u, \hat{u} \in U\,, \\ |\varphi(t, 0, u)| \le L & \forall (t, u) \in [0, T] \times U\,. \end{cases}$$

- *The maps $b,\ f$ and h are C^1 in x and there exists a modulus of continuity $\eta : [0,\infty) \to [0,\infty)$ such that,*

$$|\partial_x \varphi(t,x,u) - \partial_x \varphi(t,\hat{x},\hat{u})| \leq \eta\Big(|x-\hat{x}| + d(u,\hat{u})\Big) \quad \forall t \in [0,T]\ x,\hat{x} \in \mathbb{R}^n,\ u,\hat{u} \in U\,.$$

As discussed in [YZ99, p.102], Assumption 8.6.1 implies that (8.6.1) admits a unique solution and (8.6.2) is well defined. Let us denote by $\mathcal{U}[0,T] := \{u(\cdot) : [0,T] \to U \mid u \text{ is measurable}\}$, then optimal control problem is to find $u^* \in \mathcal{U}[0,T]$ that satisfies,

$$J(u^*) = \inf_{u \in \mathcal{U}[0,T]} J(u)\,. \tag{8.6.3}$$

Such u^* is referred to as an *optimal control*, and the corresponding $x^*(\cdot) := x(\cdot; u^*)$ the *optimal state trajectory*. We can then state the deterministic version of Pontryagin's maximum principle as [YZ99, p.103].

Theorem 8.6.2. *[Pontryagin's Maximum Principle] Let Assumption 8.6.1 hold and let (x^*,u^*) be the optimal pair to* (8.6.3). *Then, there exists a function $p : [0,T] \to \mathbb{R}^n$ satisfying the following,*

$$\begin{cases} \dot{p}(t) = -\partial_x b(t,x^*(t),u^*(t))^\intercal p(t) + \partial_x f(t,x^*(t),u^*(t)), & \textit{a.e. } t \in [0,T] \\ p(T) = -\partial_x h(x^*(T))\,, \end{cases} \tag{8.6.4}$$

and

$$H(t,x^*(t),u^*(t),p(t)) = \max_{u \in U}\{H(t,x^*(t),u,p(t))\} \quad \textit{a.e. } t \in [0,T]\,,$$

where

$$H(t,x,u,p) := \langle p, b(t,x,u)\rangle - f(t,x,u) \quad (t,x,u,p) \in [0,T] \times \mathbb{R}^n \times U \times \mathbb{R}^n\,.$$

Typically p is referred to as the *adjoint function* and (8.6.4) the *adjoint equation*, and the function H is called the *Hamiltonian*.

Remark 8.6.3 (An alternative approach). *The maximum principle is not the only way one can use to solve this problem. An alternative is by solving the so-called Hamilton-Jacobi-Bellman (HJB) equation. This approach is typically more difficult since in this case the HJB is typically a non linear first order PDE.*

Chapter 9

Simulation of McKean-Vlasov SDEs with Super-Linear Growth

The aim of this chapter is to develop a numerical scheme for simulating a McKean-Vlasov Stochastic Differential Equations (MV-SDEs) with drifts of super-linear growth and Lipschitz diffusion coefficients (with linear growth).

Similar to standard SDEs, MV-SDEs have been shown to have a unique strong solution in the super-linear growth setting in spatial parameter setting, see [dRST17]. Of course, many mean-field models exhibit non globally Lipschitz growth, for example, mean-field models for neuronal activity (e.g. stochastic mean-field FitzHugh-Nagumo models or the network of Hodgkin-Huxley neurons) [BFFT12], [BCC11], [BFT15] appearing in biology or physics [DGG^{+}11], [DFG^{+}16]. We refer to the review in [BFFT12] for further motivation of the problem.

Closer to our work, we highlight: [BF17] develop an explicit Euler scheme to deal with a specific MV-SDE type equation; convergence is given but under Lipschitz conditions and constant diffusion coefficient. [Mal03] studies an implicit Euler scheme in order to approximate a specific equation and requires constant diffusion coefficient, symmetry and uniform convexity of the interaction potential.

Our Contribution. Firstly, we show that the above particle scheme converges in the super-linear growth case without coercivity/dissipativity (propagation of chaos). This result is crucial in showing convergence of the numerical scheme to the particle system rather than to the original MV-SDE, with corresponding rate.

The second contribution is the development and strong convergence of the explicit scheme to the MV-SDE, inspired by the explicit scheme originally developed in [HJK12], [Sab13]. We also obtain the classical $1/2$ rate of convergence in the stepsize. Combining this with the propagation of chaos result gives an overall convergence rate for the explicit scheme.

The final contribution is to show strong convergence of an implicit scheme. This turns out to be a challenging problem since results involving implicit schemes rely on stopping time arguments. This causes several issues when generalising results to the MV-SDE setting and we have had to make stronger assumptions on the coefficients in this setting in order for the arguments to continue to hold. On the other hand, we allow for both random initial conditions and time dependent coefficients that to the best of our knowledge have not been fully treated in the standard SDE setting. We discuss these issues in Remarks 9.1.5 and 9.3.11. We only focus on strong convergence of this scheme and not the rate, mainly because the explicit scheme is in general superior (as our numerical testing shows) and such proof would lead to lengthy statements below without substantially enhancing the scope of our work.

From a technical point of view, we highlight the successful use of stopping time arguments in combination with McKean-Vlasov equations and associated particle systems to show the convergence of the implicit scheme.

The chapter is structured in the following way. In Section 9.1, we state our main result, namely, propagation of chaos and convergence results for the two schemes. Following that, in Section 9.2 we provide several numerical examples and highlight the *particle corruption* phenomena. This analysis implies one cannot hope to build a reliable scheme based on a standard Euler scheme. We further show the increased computational complexity associated with a MV-SDE makes the implicit scheme a less viable option than the explicit (tamed) scheme. Finally, the proofs are given in Section 9.3.

Standard Euler scheme particle system. In general one cannot simulate (8.2.2) directly and therefore turns to a numerical scheme such as Euler. We partition the time interval $[0,T]$ into M steps of size $h := T/M$, we then define $t_k := kh$ and recursively define the particle system for $k \in \{0, \dots, M-1\}$ as,

$$\bar{X}^{i,N,M}_{t_{k+1}} = \bar{X}^{i,N,M}_{t_k} + b\big(t_k, \bar{X}^{i,N,M}_{t_k}, \bar{\mu}^{X,N}_{t_k}\big)h + \sigma\big(t_k, \bar{X}^{i,N,M}_{t_k}, \bar{\mu}^{X,N}_{t_k}\big)\Delta W^i_{t_k},$$

where $\bar{\mu}^{X,N}_{t_k}(\mathrm{d}x) := \frac{1}{N}\sum_{j=1}^N \delta_{\bar{X}^{j,N,M}_{t_k}}(\mathrm{d}x)$, $\Delta W^i_{t_k} := W^i_{t_{k+1}} - W^i_{t_k}$ and $\bar{X}^{i,N,M}_0 := X^i_0$. Under Lipschitz regularity it is well known that this scheme converges, see [BT97] or [KHO97] (here a weak rate of convergence is shown under an additional regularity assumption).

Euler particle system for the super-linear case: Explicit and Implicit. However, as discussed in works such as [HJK11], [HJK12], [Sab13] one does not have convergence of the Euler scheme when we move away from the global Lipschitz setting. The goal of this chapter is to construct suitable numerical schemes that converge. Inspired by the above works we consider a so-called *tamed* Euler scheme. With the notation above consider the following scheme

$$\bar{X}^{i,N,M}_{t_{k+1}} = \bar{X}^{i,N,M}_{t_k} + \frac{b\big(t_k, \bar{X}^{i,N,M}_{t_k}, \bar{\mu}^{X,N}_{t_k}\big)}{1 + M^{-\alpha}\big|b\big(t_k, \bar{X}^{i,N,M}_{t_k}, \bar{\mu}^{X,N}_{t_k}\big)\big|}h + \sigma\big(t_k, \bar{X}^{i,N,M}_{t_k}, \bar{\mu}^{X,N}_{t_k}\big)\Delta W^i_{t_k}, \quad (9.0.1)$$

where $\bar{\mu}^{X,N}_{t_k}(\mathrm{d}x) = \frac{1}{N}\sum_{j=1}^N \delta_{\bar{X}^{j,N,M}_{t_k}}(\mathrm{d}x)$ and $\alpha \in (0, 1/2]$ with $\bar{X}^{i,N,M}_0 = X^i_0$.

Of course, explicit schemes are not the only method one can deploy to solve this problem, we also consider the following implicit scheme

$$\tilde{X}^{i,N,M}_{t_{k+1}} = \tilde{X}^{i,N,M}_{t_k} + b\big(t_k, \tilde{X}^{i,N,M}_{t_{k+1}}, \tilde{\mu}^{X,N,M}_{t_k}\big)h + \sigma\big(t_i, \tilde{X}^{i,N,M}_{t_k}, \tilde{\mu}^{X,N,M}_{t_k}\big)\Delta W^i_{t_k}, \quad (9.0.2)$$

where $\tilde{\mu}^{X,N,M}_{t_k}(\mathrm{d}x) := \frac{1}{N}\sum_{j=1}^N \delta_{\tilde{X}^{j,N,M}_{t_k}}(\mathrm{d}x)$ and $\tilde{X}^{i,N,M}_0 = X^i_0$.

Remark 9.0.1. *There are other explicit types of explicit schemes that can handle non Lipschitz growth, such as the truncation in* [CJM16]. *For this work though we shall focus on work stemming from* [HJK12].

9.1 Main Results

We state our main results and assumption here, the proofs are postponed to Section 9.3. For this work we put additional assumption on the time dependence,

Assumption 9.1.1. *Assume that b and σ are $1/2$-Hölder continuous in time.*

Recall that we want to associate a particle system to the MV-SDE and show its convergence, so-called *propagation of chaos*. We have the following result that holds under weaker assumptions than those in Theorem 9.1.3.

Proposition 9.1.2 (Propagation of chaos)**.** *Let Assumption 9.1.1 and the assumptions in Theorem 8.2.2 hold for $m > 4$. Then we have the following convergence result.*

$$\sup_{1 \le i \le N} \mathbb{E}\big[\sup_{0 \le t \le T} |X^i_t - X^{i,N}_t|^2\big] \le C \begin{cases} N^{-1/2} & \text{if } d < 4, \\ N^{-1/2}\log(N) & \text{if } d = 4, \\ N^{-2/d} & \text{if } d > 4. \end{cases}$$

Therefore, to show convergence between our numerical scheme and the MV-SDE, we only need to show that the "true" particle scheme and numerical version of the particle scheme converge.

Explicit scheme

We first introduce the continuous time version of the explicit scheme. Denote by $\kappa(t) := \sup\{s \in \{0, h, 2h, \dots, Mh\} : s \le t\}$ for all $t \in [0,T]$,

$$b_M(t,x,\nu) := \frac{b(t,x,\nu)}{1 + M^{-\alpha}|b(t,x,\nu)|}\,,$$

with $\alpha \in (0, 1/2]$ for all $t \in [0,T]$, $x \in \mathbb{R}^d$, $\nu \in \mathcal{P}_2(\mathbb{R}^d)$

$$\begin{aligned} X_t^{i,N,M} &= X_0^i + \int_0^t b_M\left(\kappa(s), X_{\kappa(s)}^{i,N,M}, \mu_{\kappa(s)}^{X,N,M}\right)\mathrm{d}s + \int_0^t \sigma\left(\kappa(s), X_{\kappa(s)}^{i,N,M}, \mu_{\kappa(s)}^{X,N,M}\right)\mathrm{d}W_s^i, \\ \mu_t^{X,N,M}(\mathrm{d}x) &= \frac{1}{N}\sum_{j=1}^N \delta_{X_t^{j,N,M}}(\mathrm{d}x). \end{aligned} \tag{9.1.1}$$

Note that $|b_M(t,x,\nu)| \le \min\left(M^\alpha, |b(t,x,\nu)|\right)$ and that $\bar{X}_{t_k}^{i,N,M} = X_{t_k}^{i,N,M}$ for all $k \in \{0, 1, \dots, M\}$ and hence $X^{i,N,M}$ is a continuous version of $\bar{X}^{i,N,M}$ from (9.0.1). This then leads to our main explicit scheme convergence result.

Theorem 9.1.3 (Strong Convergence of Explicit). *Let Assumption 8.2.1 and 9.1.1 hold, further let $X_0 \in L^m(\mathbb{R}^d)$ for $m \ge 4(1+q)$ (note $q > 1$). Let X^i be the solution to* (8.2.3)*, and $X^{i,N,M}$ be as in* (9.1.1) *with $\alpha = 1/2$. Then we obtain the following convergence result*

$$\sup_{1\le i\le N} \mathbb{E}\big[\sup_{0\le t\le T} |X_t^i - X_t^{i,N,M}|^2\big] \le C \begin{cases} N^{-1/2} + h & \text{if } d < 4, \\ N^{-1/2}\log(N) + h & \text{if } d = 4, \\ N^{-2/d} + h & \text{if } d > 4. \end{cases}$$

The following convergence result is crucial to the above theorem.

Proposition 9.1.4. *Let the assumption in Theorem 9.1.3 hold. Then it holds that*

$$\sup_{1\le i\le N} \mathbb{E}\big[\sup_{0\le t\le T} |X_t^{i,N} - X_t^{i,N,M}|^2\big] \le Ch.$$

Proof of Theorem 9.1.3. Theorem 9.1.3 is a consequence of Propositions 9.1.2 and 9.1.4. □

Remark 9.1.5 (Issues using stopping times). *The technique of using the stopping time $\tau_R^i := \inf\{t \ge 0 : |X_t^{i,N,M}| \ge R\}$ to control the particles is suboptimal and several problems appear by introducing them. Namely, one can only consider stopping times that stop one particle since otherwise the convergence speed would decrease with a higher number of particles. However, applying a stopping time to a single particle does not allow us to fully bound the coefficients and moreover destroys the result of all particles being identically distributed.*

The stopping times arguments used for the implicit scheme below require stronger assumptions in order to make the theory hold.

Implicit scheme

We have shown convergence of the explicit scheme for non Lipschitz coefficients, although this is indeed not the only method, there is another popular method known as implicit or backward Euler scheme. That being said, the implicit scheme has some well documented disadvantages, namely it is expensive compared to its explicit counterpart, we discuss this issue further in Section 9.2. One can consult, [MS13] for example on the implicit scheme (and extensions) for standard SDEs.

Standard implicit scheme convergence results rely on the so called monotone growth condition, we therefore proceed with the following assumption.

Assumption 9.1.6.

(H1). There exists a constant C such that, for all $\mu \in \mathcal{P}_2(\mathbb{R}^d)$,

$$|b(0,0,\mu)| + |\sigma(0,0,\mu)| \leq C\,.$$

(H2). σ is only a function of time and space (does not have a measure dependence).

Although the main convergence theorem requires both H1 and H2, we only use H2 at the end of the proof of convergence. We present our auxiliary results requiring only H1 as we believe them to be of general independent interest.

Remark 9.1.7 (Monotone Growth)**.** *The combination of Assumption 8.2.1, 9.1.1 and H1, imply the monotone growth condition. Namely, there exist constants α and β such $\forall\, t \in [0,T], \mu \in \mathcal{P}_2(\mathbb{R}^d)$ with l being the dimension of the BM,*

$$\langle x, b(t,x,\mu)\rangle + \frac{1}{2}\sum_{a=1}^{l} |\sigma_a(t,x,\mu)|^2 \leq \alpha + \beta|x|^2.$$

We now state the strong convergence of the implicit scheme (9.0.2) to (8.2.2).

Proposition 9.1.8. *Let Assumption 8.2.1, 9.1.1 and 9.1.6 hold. Fix a timestep $h^* < 1/\max(L_b, 2\beta)$ and assume $X_0 \in L^{4(q+1)}(\mathbb{R}^d)$. Then, for any $T = Mh$ and $s \in [1,2)$*

$$\sup_{1\leq i\leq N} \lim_{h\to 0} \mathbb{E}[|X_T^{i,N} - \tilde{X}_T^{i,N,M}|^s] = 0\,.$$

Theorem 9.1.9 (Strong Convergence of Implicit Scheme)**.** *Let the Assumption in Proposition 9.1.8 hold. Then, for any $T = Mh$ and $s \in [1,2)$ one has*

$$\lim_{N\to\infty} \sup_{1\leq i\leq N} \lim_{h\to 0} \mathbb{E}[|X_T^{i} - \tilde{X}_T^{i,N,M}|^s] = 0\,.$$

Proof. The proof of this result follows by combing Proposition 9.1.2 and 9.1.8 and noting that the assertion in Proposition 9.1.8 is independent of N. □

9.1.1 Further Discussion of Results

The results presented here are focused on the simulation of MV-SDEs with superlinear coefficients. This widens the application of MV-SDEs considerably since many applications require non Lipschitz growth. That being said we are still making a rather strong assumption on the measure dependence and we do not allow for general cross terms e.g. $X\mathbb{E}[X]$. The main reason for this is, at present, there is no general theory regarding existence and uniqueness of MV-SDEs with such general coefficients. Consequently one must understand these more fundamental results before constructing numerical schemes. However, there is a vast amount of research going into MV-SDEs and in the near future such existence and uniqueness results may be proved. In that case the numerical schemes above may be able to handle these more general equations or may require a further slight modification.

It is our belief that Assumption 9.1.6 although sufficient, is not necessary to guarantee the implicit scheme converges. As research is carried out into stopping times and MV-SDEs, future theoretical developments in this direction may allow this assumption to be weakened.

9.2 Numerical testing and Examples

We illustrate immediately our results with numerical examples. We highlight the issues of using the standard Euler scheme in this setting and also compare the computational time and complexity of the explicit and implicit scheme. We juxtapose our findings to those in [BFFT12].

9.2.1 Particle Corruption

It is well known that the Euler scheme fails (diverges) when one moves outside the realm of linear growing coefficients, see [HJK11]. We claim that this divergence is worse in the setting of MV-SDEs and associated particle system due to an effect we refer to as *particle corruption*.

The basic idea is that one particle becomes influential on all other particles, thus we are no longer in the setting of "weakly interacting". This is of course not a problem for standard SDE simulation. We show two aspects of particle corruption in a simple example, firstly it exists i.e. one particle can cause the whole system to crash. Secondly and perhaps more profoundly, the more particles one has the more likely this is. This is of course a devastating issue when simulating a MV-SDE since accurately approximating the measure depends on having a large number of interacting particles.

To show this example we take a classical non-globally Lipschitz SDE, the stochastic Ginzburg Landau equation (see [Tie13]) and add a simple mean field term to it,

$$\mathrm{d}X_t = \Big(\frac{\sigma^2}{2}X_t - X_t^3 + c\mathbb{E}[X_t]\Big)\mathrm{d}t + \sigma X_t \mathrm{d}W_t, \quad X_0 = x.$$

This MV-SDE clearly satisfies the assumption to have a unique strong solution in $\mathbb{S}^p$ for all $p > 1$, hence in theory one could calculate $\varphi(t) := \mathbb{E}[X_t]$ and have a standard SDE with one-sided Lipschitz drift. The analysis carried out in [HJK11] then implies that the Euler scheme diverges here.

Showing particle corruption exists. For our example we simulate $N = 5000$ particles with a time step $h = 0.05$, $T = 2$ and $X_0 = 1$, we also take $\sigma = 3/2$ and $c = 1/2$. We rerun this example until we observed a blow up and plotted the particle paths in Figure 9.1.

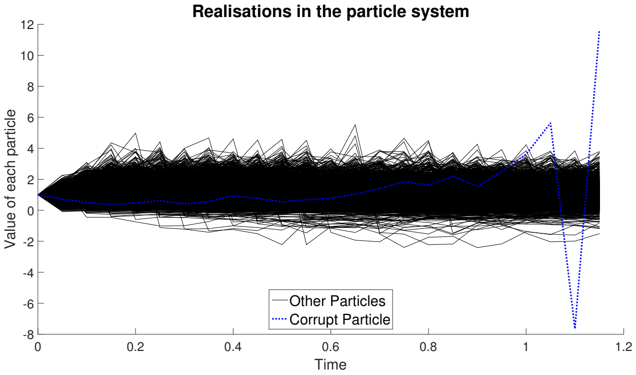


Figure 9.1: Showing the realisations of the particles in the system. We note that the particle given by the dashed line is starting to oscillate and is taking larger values than its surrounding particles.

Figure 9.1 show the first part of the divergence, namely all particles are reasonably well behaved until one starts to oscillate rapidly. We have stopped plotting before the time boundary since this particle diverges shortly after this. We refer to this particle as the *corrupt particle* and it is fairly straightforward to see it will diverge. However, due to the interaction this single particle influences all the remaining particles and the whole system diverges shortly after.

Remark 9.2.1 (Why is particle corruption so pronounced?)**.** *The reason this effect is so dramatic is a simple consequence of the mean-field interaction. Typically, one observes divergence of the Euler scheme via a handful of Monte Carlo simulations that return extremely large (or infinite) values. When one then looks to calculate the expected value of the SDEs at the terminal time for example, these few events completely dominate the other results. This is summed up in a statement of* [HJK11], *where an exponentially small probability event has a double exponential impact.*

The difference in the MV-SDE (weakly interacting particle) case is that the expectation appears inside the simulation, hence a divergence of a single particle influences multiple particles simultaneously during the simulation and not just at the final time.

Convergence of Euler and propagation of chaos is impossible. The above shows that one particle diverging can cause the whole system to diverge, one may argue that using more particles would reduce the dependency between them and hence influence the system less. In fact as we shall see the opposite is true, the more particles the more likely a divergence is. To test this we use the same example as above but use $N = [1000,\ 5000,\ 10000,\ 20000]$ particles and rerun each case 1000 times and record the total number of times we observe a divergence over the ensemble.

Number of particles	1000	5000	10000	20000
Number of blow ups	3	32	43	108

Table 9.1: Number of divergences recorded at each particle level out of 1000 simulations.

The results in Table 9.1 show conclusively that the more particles the more likely a divergence is to occur. This is a real problem in this setting since in order to minimise the propagation of chaos error one should take N as large as possible, but in doing so makes the Euler scheme approximation (likelier to) diverge.

Remark 9.2.2 (Euler cannot work). *We have shown that naively applying the standard Euler scheme in the MV-SDE setting with non globally Lipschitz coefficient has issues. However, for standard SDEs there are some simple fixes one can apply and still obtain convergence e.g. removing paths that leave some ball as considered in* [MT05]. *Methods like this cannot work here since, we either take the ball "small" and therefore our approximation to the law is poor. Or we take a large ball, but then as the particles head towards the boundary they can "drag" other particles with them which again makes the system unstable.*

The dependence on the measure (other particles) implies that the more crude approximation techniques cannot yield the strong convergence results we obtain with the more sophisticated techniques presented here. In [BFFT12] *the authors have a non-globally Lipschitz MV-SDE and simulate using standard Euler scheme. Since no divergence was observed in their simulations they conjectured that the Euler scheme works in their setting, however, they used a "small" diffusion coefficient ($\sigma \in [0, 0.5]$) and small particle number (in the order of hundreds), which makes divergence unlikely to be observed (but not impossible) and yields poorer approximation results. Again, our methods provide certainty in terms of convergence (and convergence rate).*

9.2.2 Timing of Implicit vs Explicit: Size of cloud and spatial dimension

It is well documented that implicit schemes are slower than explicit ones, mainly because one must solve a fixed point equation at each step. This operation is not "cheap" and moreover scales d^2 in dimension, see [HJK12]. Of course this analysis is carried out for standard SDEs, what we wish to consider is how the particle system affects the timing of both methods.

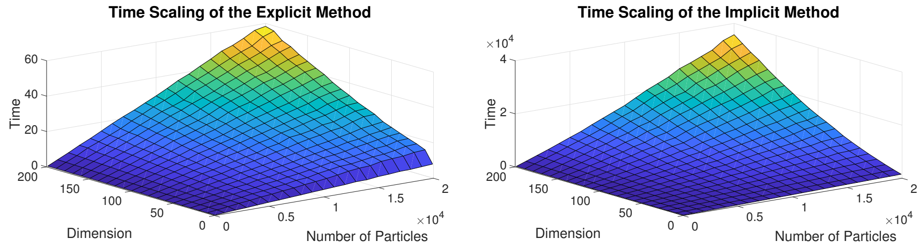
We consider the same example as previous (but take $T = 1$), we then consider a set of dimensions from 1 to 200 and number of particles from 100 to 20000. Plotting the time taken for both methods is given in Figure 9.2.

Firstly, we observe that the explicit scheme is two to three orders of magnitude faster than the implicit scheme. At the highest dimensional and particle number this difference is very apparent with the tamed scheme taking approximately 1 minute and the implicit 10 hours. Another note to make is the scaling of each method, both methods scale similarly with particle number , but the tamed scheme scales linearly with dimension, this is superior to the d^2 scaling of the implicit scheme.

Even for the case $d = 1$, $N = 20000$ the tamed scheme takes approximately 7 seconds while the implicit scheme takes approximately 23 minutes. For many practical applications $N = 20000$ is not enough for an acceptable level of accuracy, with this in mind and the dimension scaling, this makes the implicit scheme a very expensive method in this setting.

9.2.3 Explicit Vs Implicit Convergence: the Neuron Network Model

We compare the convergence of the explicit and the implicit scheme. To this end we use the system in [BFFT12] where the authors develop a non globally Lipschitz MV-SDE to model neuron



(a) Explicit Scheme (b) Implicit Scheme

Figure 9.2: Showing how the time (in seconds) of the explicit scheme (left; timescale $\approx$ 60 seconds) and implicit scheme (right; timescale $\approx 10^4$ seconds) changes with particles and dimension.

activity. In our notation their system with $b : [0,T]\times\mathbb{R}^3\times\mathcal{P}_2(\mathbb{R}^3) \to \mathbb{R}^3$, $\sigma : [0,T]\times\mathbb{R}^3\times\mathcal{P}_2(\mathbb{R}^3) \to \mathbb{R}^{3\times 3}$ reads for $x = (x_1, x_2, x_3), z = (z_1, z_2, z_3) \in \mathbb{R}^3$ as

$$b\left(t,x,\mu\right) := \begin{pmatrix} x_1 - (x_1)^3/3 - x_2 + I - \int_{\mathbb{R}^3} J\left(x_1 - V_{rev}\right) z_3 \mathrm{d}\mu(z) \\ c\left(x_1 + a - bx_2\right) \\ a_r \frac{T_{max}(1-x_3)}{1+\exp(-\lambda(x_1-V_T))} - a_d x_3 \end{pmatrix}$$

$$\sigma\left(t,x,\mu\right) := \begin{pmatrix} \sigma_{ext} & 0 & -\int_{\mathbb{R}^3} \sigma_J\left(x_1 - V_{rev}\right) z_3 \mathrm{d}\mu(z) \\ 0 & 0 & 0 \\ 0 & \sigma_{32}(x) & 0 \end{pmatrix}$$

with

$$\sigma_{32}(x) := \mathbb{1}_{\{x_3\in(0,1)\}} \sqrt{a_r \frac{T_{max}(1-x_3)}{1+\exp(-\lambda(x_1-V_T))} + a_d x_3}\, \Gamma \exp(-\Lambda/(1-(2x_3-1)^2)),$$

$T = 2$ is chosen as the final time and

$$X_0 \sim \mathcal{N}\left(\begin{pmatrix} V_0 \\ w_0 \\ y_0 \end{pmatrix}, \begin{pmatrix} \sigma_{V_0} & 0 & 0 \\ 0 & \sigma_{w_0} & 0 \\ 0 & 0 & \sigma_{y_0} \end{pmatrix} \right),$$

where the parameters have the values

$$\begin{array}{ccccccc} V_0 = 0 & \sigma_{V_0} = 0.4 & a = 0.7 & b = 0.8 & c = 0.08 & I = 0.5 & \sigma_{ext} = 0.5 \\ w_0 = 0.5 & \sigma_{w_0} = 0.4 & V_{rev} = 1 & a_r = 1 & a_d = 1 & T_{max} = 1 & \lambda = 0.2 \\ y_0 = 0.3 & \sigma_{y_0} = 0.05 & J = 1 & \sigma_J = 0.2 & V_T = 2 & \Gamma = 0.1 & \Lambda = 0.5. \end{array}$$

As the true solution is unknown to compare the convergence rates, we use as proxy the output of the explicit scheme with 2^{23} steps. Since the explicit scheme has convergence rate $\sqrt{h}$ we know that 2^{16} steps and below yields one order of magnitude larger errors. The simulation for 1000 particles and average root mean square error of each particle is given in Figure 9.3.

One can observe that although initially the implicit scheme has a better rate of convergence, it levels off to yield the expected $1/2$ rate. Making the explicit scheme the more computationally efficient. Of course our “true” was calculated from the explicit scheme, hence we additionally carried out a similar test with a “true” from the implicit, and the results were almost identical.

Remark 9.2.3 (Small Diffusion Setting)**.** *Above, we have taken $\sigma_{ext} = 0.5$, this goes against the example in* [BFFT12] *where $\sigma_{ext} = 0$. As it turns out, in the case $\sigma_{ext} = 0$, the implicit scheme has a convergence rate close to 1 (up to an error of around 10^{-4}), while the explicit scheme maintains the standard $1/2$ rate. It is our belief that this is due to the fact that when $\sigma_{ext} = 0$ the diffusion coefficient makes little difference, hence both scheme revert close to their deterministic convergence rate. The explicit scheme of course still rate of order $1/2$, while the implicit is order 1. It may therefore be that in the setting of small diffusion terms the implicit can yield superior results, of*

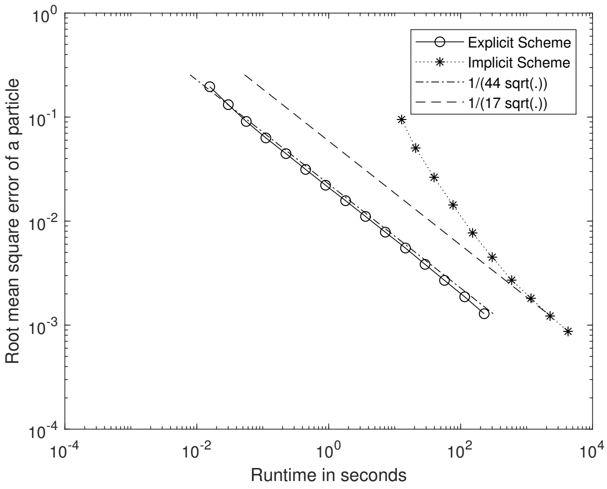


Figure 9.3: Root mean square error of the explicit and implicit. The number of steps of the explicit scheme are $M \in \{2^2, 2^3, \dots, 2^{16}\}$ and of the implicit scheme are $M \in \{2^2, 2^3, \dots, 2^{11}\}$. We used 1000 particles and the true is calculated from the explicit with 2^{23} steps. Both schemes converge with rate $1/2$.

course though this is a special case and is not true in general.

Obtaining the Density

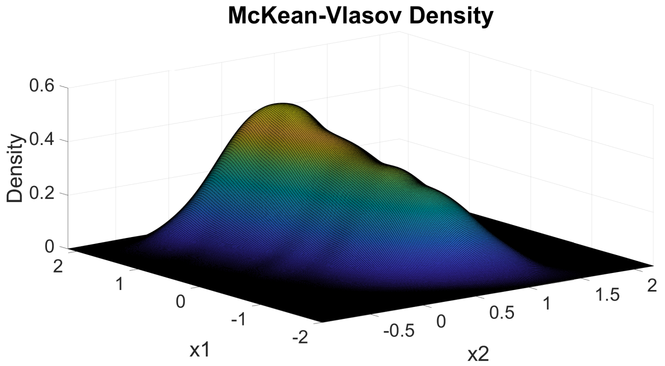


Figure 9.4: Approximate density of the first and second component of the MV-SDE at time $T = 1.2$. We used 10000 particles, 2^{20} steps and a bandwidth of 0.15 in the kernel smoothing.

In some applications as well as the value of the MV-SDE at the terminal time, one may also be interested in the density (law). In [BFFT12, Section 4] the authors compare density estimation using both the Fokker-Plank equation and the histogram from the particle system. The approach using PDEs becomes computationally expensive here if one considers multiple populations of MV-SDE and hence the authors take a simple case (see [BFFT12, Section 4.3]). There are of course other drawbacks such as dimension scaling which often make stochastic techniques more favorable in this setting. Moreover, using the PDE one will only obtain the density, if one is further interested in calculating a “payoff” i.e. $\mathbb{E}[G(X_T)]$ for some function G. Then we would require an additional integral approximation or Metropolis Hastings style sampling scheme to calculate this expectation. While [BFFT12] apply a basic histogram approach when using MV-SDEs, this does not yield particularly nice results, namely, the resultant density is not a smooth surface. There are however, many statistical techniques one can use to improve this, see [Kee11, Chapter 18.4] for further results and discussion. Taking the example in [BFFT12] (with $\sigma_{ext} = 0$) and applying MATLAB’s `ksdensity` function we obtain Figure 9.4.

One can observe the similarity between our result using SDEs and the one obtained in [BFFT12, pg 31] using the (expensive) PDE approach.

9.3 Proof of Main Results

We shall use C to denote a constant that can changes from line to line, but only depend on known quantities, T, d, the one-sided Lipschitz coefficients etc.

9.3.1 Propagation of Chaos

Let us show the propagation of chaos result.

Proposition 9.1.2. Let $t \in [0,T]$ and fix $1 \leq i \leq N$, we then approach the proof in the usual way for dealing with one-sided Lipschitz coefficients, namely we apply Itô's formula to the difference (note X_0^i cancel),

$$\begin{aligned}|X_t^i - X_t^{i,N}|^2 = &\int_0^t 2\langle X_s^i - X_s^{i,N}, b(s,X_s^i,\mu_s) - b(s,X_s^{i,N},\bar{\mu}_s^N)\rangle \mathrm{d}s \\ &+ \int_0^t 2\langle X_s^i - X_s^{i,N}, (\sigma(s,X_s^i,\mu_s) - \sigma(s,X_s^{i,N},\bar{\mu}_s^N))\mathrm{d}W_s^i\rangle \\ &+ \sum_{a=1}^{l} \int_0^t |\sigma_a(s,X_s^i,\mu_s) - \sigma_a(s,X_s^{i,N},\bar{\mu}_s^N)|^2 \mathrm{d}s\,, \end{aligned} \tag{9.3.1}$$

where σ_a is the ath column of matrix σ, hence σ_a is a d-dimensional vector. Considering the first integral in (9.3.1),

$$\begin{aligned}\langle X_s^i - X_s^{i,N}, b(s,X_s^i,\mu_s) - b(s,X_s^{i,N},\bar{\mu}_s^N)\rangle =& \langle X_s^i - X_s^{i,N}, b(s,X_s^i,\mu_s) - b(s,X_s^{i,N},\mu_s)\rangle \\ &+ \langle X_s^i - X_s^{i,N}, b(s,X_s^{i,N},\mu_s) - b(s,X_s^{i,N},\bar{\mu}_s^N)\rangle.\end{aligned}$$

Applying the one-sided Lipschitz property in space and $W^{(2)}$ in measure along with Cauchy-Schwarz we obtain,

$$\langle X_s^i - X_s^{i,N}, b(s,X_s^i,\mu_s) - b(s,X_s^{i,N},\bar{\mu}_s^N)\rangle \leq C|X_s^i - X_s^{i,N}|^2 + C|X_s^i - X_s^{i,N}|W^{(2)}(\mu_s,\bar{\mu}_s^N)\,.$$

As is done in [Car16], we introduce the empirical measure constructed from the true solution i.e. $\mu_s^N := \frac{1}{N}\sum_{j=1}^N \delta_{X_s^j}$. Since $W^{(2)}$ is a metric (see [Vil08, Chapter 6]), we have

$$W^{(2)}(\mu_s,\bar{\mu}_s^N) \leq W^{(2)}(\mu_s,\mu_s^N) + W^{(2)}(\mu_s^N,\bar{\mu}_s^N)\,.$$

Since μ_s^N, $\bar{\mu}_s^N$ are empirical measures a standard result for Wasserstein metric is

$$W^{(2)}(\mu_s^N,\bar{\mu}_s^N) \leq \Big(\frac{1}{N}\sum_{j=1}^N |X_s^j - X_s^{j,N}|^2\Big)^{1/2}.$$

We leave the other $W^{(2)}$ term for the moment and consider the diffusion coefficient in the time integral. Since σ is globally Lipschitz and $W^{(2)}$ for each a (by definition $\sigma_a = \sigma e_a$, with e_a the basis vector, global Lipschitz follows from our norm).

$$\begin{aligned}&|\sigma_a(s,X_s^i,\mu_s) - \sigma_a(s,X_s^{i,N},\bar{\mu}_s^N)|^2 \\ &\leq C\big(|\sigma_a(s,X_s^i,\mu_s) - \sigma_a(s,X_s^{i,N},\mu_s)|^2 + |\sigma_a(s,X_s^{i,N},\mu_s) - \sigma_a(s,X_s^{i,N},\bar{\mu}_s^N)|^2\big) \\ &\leq C\big(|X_s^i - X_s^{i,N}|^2 + W^{(2)}(\mu_s,\bar{\mu}_s^N)^2\big) \\ &\leq C\big(|X_s^i - X_s^{i,N}|^2 + \frac{1}{N}\sum_{j=1}^N |X_s^j - X_s^{j,N}|^2 + W^{(2)}(\mu_s,\mu_s^N)^2\big)\,.\end{aligned}$$

One can note this is independent of a. The final term to bound is the stochastic integral term, to do this though we take supremum and expectation to (9.3.1)

$$\begin{aligned}
&\mathbb{E}\Big[\sup_{t\in[0,T]}|X_t^i - X_t^{i,N}|^2\Big] \\
&\quad\le C\mathbb{E}\Big[\sup_{t\in[0,T]}\int_0^t |X_s^i - X_s^{i,N}|^2 + |X_s^i - X_s^{i,N}|W^{(2)}(\mu_s,\bar{\mu}_s^N)\mathrm{d}s\Big] \\
&\quad+\mathbb{E}\Big[\sup_{t\in[0,T]}\int_0^t 2\langle X_s^i - X_s^{i,N}, (\sigma(s,X_s^i,\mu_s)-\sigma(s,X_s^{i,N},\bar{\mu}_s^N))\mathrm{d}W_s^i\rangle\Big] \\
&\quad+Cl\mathbb{E}\Big[\sup_{t\in[0,T]}\int_0^t |X_s^i - X_s^{i,N}|^2 + \frac{1}{N}\sum_{j=1}^N |X_s^j - X_s^{j,N}|^2 + W^{(2)}(\mu_s,\bar{\mu}_s^N)^2\mathrm{d}s\Big]\,. \qquad (9.3.2)
\end{aligned}$$

For the stochastic integral,

$$\begin{aligned}
&\mathbb{E}\Big[\sup_{t\in[0,T]}\int_0^t 2\langle X_s^i - X_s^{i,N}, (\sigma(s,X_s^i,\mu_s)-\sigma(s,X_s^{i,N},\bar{\mu}_s^N))\mathrm{d}W_s^i\rangle\Big] \\
&\quad\le\mathbb{E}\Big[\sup_{t\in[0,T]}\Big|\int_0^t 2\langle X_s^i - X_s^{i,N}, (\sigma(s,X_s^i,\mu_s)-\sigma(s,X_s^{i,N},\bar{\mu}_s^N))\mathrm{d}W_s^i\rangle\Big|\Big] \\
&\quad\le C\mathbb{E}\Big[\Big(\int_0^T\Big(\sum_{a=1}^l |\sigma_a(s,X_s^i,\mu_s)-\sigma_a(s,X_s^{i,N},\bar{\mu}_s^N)|^2\Big)|X_s^i - X_s^{i,N}|^2\mathrm{d}s\Big)^{1/2}\Big] \\
&\quad\le\mathbb{E}\Big[\Big(\sup_{t\in[0,T]}|X_t^i - X_t^{i,N}|^2 C\int_0^T\sum_{a=1}^l |\sigma_a(s,X_s^i,\mu_s)-\sigma_a(s,X_s^{i,N},\bar{\mu}_s^N)|^2\mathrm{d}s\Big)^{1/2}\Big]\,,
\end{aligned}$$

where we have applied Burkholder-Davis-Gundy to remove the stochastic integral. Using Young's inequality $ab \le a^2/2 + b^2/2$ we can bound this term by,

$$\mathbb{E}\Big[\frac{1}{2}\sup_{t\in[0,T]}|X_t^i - X_t^{i,N}|^2 + \frac{C}{2}\int_0^T\sum_{a=1}^l |\sigma_a(s,X_s^i,\mu_s)-\sigma_a(s,X_s^{i,N},\bar{\mu}_s^N)|^2\mathrm{d}s\Big]\,.$$

Substituting into (9.3.2) yields,

$$\begin{aligned}
&\mathbb{E}\Big[\sup_{t\in[0,T]}|X_t^i - X_t^{i,N}|^2\Big] \\
&\quad\le C\mathbb{E}\Big[\sup_{t\in[0,T]}\int_0^t |X_s^i - X_s^{i,N}|^2 + |X_s^i - X_s^{i,N}|W^{(2)}(\mu_s,\bar{\mu}_s^N)\mathrm{d}s\Big] \\
&\quad+\mathbb{E}\Big[\frac{1}{2}\sup_{t\in[0,T]}|X_t^i - X_t^{i,N}|^2 + \frac{C}{2}\int_0^T\sum_{a=1}^l |\sigma_a(s,X_s^i,\mu_s)-\sigma_a(s,X_s^{i,N},\bar{\mu}_s^N)|^2\mathrm{d}s\Big] \\
&\quad+C\mathbb{E}\Big[\sup_{t\in[0,T]}\int_0^t |X_s^i - X_s^{i,N}|^2 + \frac{1}{N}\sum_{j=1}^N |X_s^j - X_s^{j,N}|^2 + W^{(2)}(\mu_s,\bar{\mu}_s^N)^2\mathrm{d}s\Big]\,.
\end{aligned}$$

Taking the $\frac{1}{2}\sup_{t\in[0,T]}|X_t^i - X_t^{i,N}|^2$ to the other side, noting that the supremum value over the integrals is $t = T$ and using the bound for the difference in σ we obtain,

$$\begin{aligned}
\mathbb{E}\Big[\sup_{t\in[0,T]}|X_t^i - X_t^{i,N}|^2\Big] &\le C\mathbb{E}\Big[\int_0^T |X_s^i - X_s^{i,N}|^2 + |X_s^i - X_s^{i,N}|W^{(2)}(\mu_s,\bar{\mu}_s^N)\mathrm{d}s\Big] \\
&\quad+C\mathbb{E}\Big[\int_0^T |X_s^i - X_s^{i,N}|^2 + \frac{1}{N}\sum_{j=1}^N |X_s^j - X_s^{j,N}|^2 + W^{(2)}(\mu_s,\bar{\mu}_s^N)^2\mathrm{d}s\Big]\,.
\end{aligned}$$

To deal with the summation term, observe that since all j are identically distributed,

$$\mathbb{E}\Big[\frac{1}{N}\sum_{j=1}^{N}|X_s^j - X_s^{j,N}|^2\Big] = \mathbb{E}\big[|X_s^i - X_s^{i,N}|^2\big]\,.$$

Therefore, applying Young's inequality to $|X_s^i - X_s^{i,N}|W^{(2)}(\mu_s, \bar{\mu}_s^N)$ and taking supremum over i,

$$\begin{aligned}\sup_{1\le i\le N}\mathbb{E}\Big[\sup_{t\in[0,T]}|X_t^i - X_t^{i,N}|^2\Big] \le& C\int_0^T \sup_{1\le i\le N}\mathbb{E}\left[|X_s^i - X_s^{i,N}|^2\right] + \mathbb{E}[W^{(2)}(\mu_s, \bar{\mu}_s^N)^2]\mathrm{d}s\\ \le& C\int_0^T \mathbb{E}\left[W^{(2)}(\mu_s, \bar{\mu}_s^N)^2\right]\mathrm{d}s\,,\end{aligned}$$

where the final step follows from Grönwall's inequality. At this point, one could conclude a pathwise propagation of chaos result, see [Car16, Lemma 1.9], however, here we are interested in the rate of convergence. This is well understood for $W^{(2)}$. We use the improved version [CD17a, Theorem 5.8] of the classical convergence result [RR98, Chapter 10.2]. Provided $X_\cdot^i \in L_\cdot^p(\mathbb{R}^d)$ for any $p > 4$, which follows from [dRST17, Theorem 3.3] then for any s,

$$\mathbb{E}\left[W^{(2)}(\mu_s, \bar{\mu}_s^N)^2\right] \le C\begin{cases}N^{-1/2} & \text{if } d < 4,\\ N^{-1/2}\log(N) & \text{if } d = 4,\\ N^{-2/d} & \text{if } d > 4.\end{cases}$$

Using the result in Theorem 8.2.2 with our assumption then completes the proof. □

9.3.2 Proof of Explicit Convergence

We detail the results to prove Proposition 9.1.4. To keep expressions as compact as possible for $s \in [0, T]$ we introduce the time discretisation error,

$$\Delta X_s^{i,N,M} := X_s^{i,N} - X_s^{i,N,M}\,,$$

where $X_s^{i,N,M}$ is defined in (9.1.1). Further we use throughout the following result,

$$\mathbb{E}\Big[\frac{1}{N}\sum_{j=1}^{N}|\Delta X_s^{j,N,M}|^2\Big] = \mathbb{E}\left[|\Delta X_s^{i,N,M}|^2\right] = \sup_{1\le j\le N}\mathbb{E}\left[|\Delta X_s^{j,N,M}|^2\right],$$

which holds since every i is identically distributed.

Remark 9.3.1. *Note that for any fixed $M \ge 1$, the drift is a bounded function, moreover, σ is at most linear growth. This ensures that the explicit scheme satisfies the bound,*

$$\sup_{1\le i\le N}\sup_{0\le t\le T}\mathbb{E}[|X_t^{i,N,M}|^p] \le C(M, p, \mathbb{E}[|X_0^i|^p])\,,$$

which is finite provided $\mathbb{E}[|X_0^i|^p] < \infty$.

Lemma 9.3.2. *Suppose Assumption 8.2.1 and 9.1.1 are fulfilled and $X_0 \in L^2(\mathbb{R}^d)$, then there exists a constant C which is independent of N such that*

$$\sup_{M\ge 1}\sup_{1\le i\le N}\sup_{0\le t\le T}\mathbb{E}\left[|X_t^{i,N,M}|^2\right] < C.$$

Proof. Applying Itô's formula and restructuring to a more useful form yields,

$$\begin{aligned}\left|X_t^{i,N,M}\right|^2 = &\left|X_0^i\right|^2 + \int_0^t 2\langle X_{\kappa(s)}^{i,N,M}, b_M\left(\kappa(s), X_{\kappa(s)}^{i,N,M}, \mu_{\kappa(s)}^{X,N,M}\right)\rangle \\ &+ \sum_{a=1}^{l}\left|\sigma_a\left(\kappa(s), X_{\kappa(s)}^{i,N,M}, \mu_{\kappa(s)}^{X,N,M}\right)\right|^2 \mathrm{d}s \\ &+ \int_0^t 2\langle X_s^{i,N,M}, \sigma\left(\kappa(s), X_{\kappa(s)}^{i,N,M}, \mu_{\kappa(s)}^{X,N,M}\right)\mathrm{d}W_s^i\rangle \\ &+ \int_0^t 2\langle X_s^{i,N,M} - X_{\kappa(s)}^{i,N,M}, b_M\left(\kappa(s), X_{\kappa(s)}^{i,N,M}, \mu_{\kappa(s)}^{X,N,M}\right)\rangle \mathrm{d}s.\end{aligned}$$

Considering the final term, observe that

$$\begin{aligned}&\left|\mathbb{E}\left[\int_0^t \langle X_s^{i,N,M} - X_{\kappa(s)}^{i,N,M}, b_M\left(\kappa(s), X_{\kappa(s)}^{i,N,M}, \mu_{\kappa(s)}^{X,N,M}\right)\rangle \mathrm{d}s\right]\right| \\ &\leq \Big|\mathbb{E}\Big[\int_0^t \Big\langle \int_{\kappa(s)}^s b_M\left(\kappa(r), X_{\kappa(r)}^{i,N,M}, \mu_{\kappa(r)}^{X,N,M}\right)\mathrm{d}r + \int_{\kappa(s)}^s \sigma\left(\kappa(r), X_{\kappa(r)}^{i,N,M}, \mu_{\kappa(r)}^{X,N,M}\right)\mathrm{d}W_r^i, \\ &\qquad b_M\left(\kappa(s), X_{\kappa(s)}^{i,N,M}, \mu_{\kappa(s)}^{X,N,M}\right)\Big\rangle \mathrm{d}s\Big]\Big| \\ &\leq \Big|\sum_{k=0}^{M-1}\int_{t_k}^{t_{k+1}} \mathbb{1}_{\{s\leq t\}}\mathbb{E}\Big[\Big\langle \mathbb{E}\Big[\int_{t_k}^s b_M\Big(\kappa(r), X_{\kappa(r)}^{i,N,M}, \mu_{\kappa(r)}^{X,N,M}\Big)\mathrm{d}r \\ &\qquad + \int_{t_k}^s \sigma\Big(\kappa(r), X_{\kappa(r)}^{i,N,M}, \mu_{\kappa(r)}^{X,N,M}\Big)\mathrm{d}W_r^i\Big|\mathcal{F}_{t_k}\Big], b_M\Big(\kappa(s), X_{\kappa(s)}^{i,N,M}, \mu_{\kappa(s)}^{X,N,M}\Big)\Big\rangle\Big]\mathrm{d}s\Big| \\ &\leq \mathbb{E}\left[\int_0^t \left|b_M\left(\kappa(s), X_{\kappa(s)}^{i,N,M}, \mu_{\kappa(s)}^{X,N,M}\right)\right| \int_{\kappa(s)}^s \left|b_M\left(\kappa(r), X_{\kappa(r)}^{i,N,M}, \mu_{\kappa(r)}^{X,N,M}\right)\right| \mathrm{d}r\ \mathrm{d}s\right] \\ &\leq \int_0^t M^\alpha \int_{\kappa(s)}^s M^\alpha \mathrm{d}r\mathrm{d}s \\ &\leq tM^{2\alpha-1} \leq t,\end{aligned}$$

where we have used that $b_M(\kappa(s), X_{\kappa(s)}^{i,N,M}, \mu_{\kappa(s)}^{X,N,M})$ is $\mathcal{F}_{t_k}$-measurable for $s < t_{k+1}$ and $\alpha \leq 1/2$. Putting this together and using Assumption 8.2.1 and 9.1.1 and the bound to remove the stochastic integral we obtain

$$\begin{aligned}\mathbb{E}[|X_t^{i,N,M}|^2] &\leq \mathbb{E}[|X_0^i|^2] + C\Big(1 + \mathbb{E}\Big[\int_0^t |X_{\kappa(s)}^{i,N,M}|^2 + \frac{1}{N}\sum_{j=1}^N |X_{\kappa(s)}^{j,N,M}|^2 \mathrm{d}s\Big]\Big) \\ &\leq \mathbb{E}[|X_0^i|^2] + C\Big(1 + \int_0^t \sup_{0\leq u\leq s} \mathbb{E}\Big[|X_u^{i,N,M}|^2 + \frac{1}{N}\sum_{j=1}^N |X_u^{j,N,M}|^2\Big]\mathrm{d}s\Big),\end{aligned}$$

which furthermore yields

$$\sup_{1\leq i\leq N}\sup_{0\leq u\leq t} \mathbb{E}\left[|X_u^{i,N,M}|^2\right] \leq C\left(1 + \mathbb{E}\left[|X_0|^2\right] + \int_0^t \sup_{1\leq i\leq N}\sup_{0\leq u\leq s} \mathbb{E}\left[|X_u^{i,N,M}|^2\right]\mathrm{d}s\right) < \infty,$$

and hence by Grönwall's lemma

$$\sup_{1\leq i\leq N}\sup_{0\leq u\leq t} \mathbb{E}\left[|X_u^{i,N,M}|^2\right] < C,$$

where C is a constant which is independent of N and M. □

Lemma 9.3.3. *If Assumption 8.2.1 and 9.1.1 are fulfilled and* $X_0 \in L^2(\mathbb{R}^d)$*, then for all* $p \in (0, 2]$

we have

$$\sup_{1\leq i\leq N}\sup_{0\leq t\leq T}\mathbb{E}\left[\left|X_t^{i,N,M}-X_{\kappa(t)}^{i,N,M}\right|^p\right]\leq CM^{-p/2}, \tag{9.3.3}$$

and

$$\sup_{1\leq i\leq N}\sup_{0\leq t\leq T}\mathbb{E}\left[\left|X_t^{i,N,M}-X_{\kappa(t)}^{i,N,M}\right|^p\left|b_M\left(\kappa(t),X_{\kappa(t)}^{i,N,M},\mu_{\kappa(t)}^{X,N,M}\right)\right|^p\right]\leq C, \tag{9.3.4}$$

where C is a positive constant independent of N and M. Furthermore, if for $p>2$

$$\sup_{M\geq 1}\sup_{1\leq i\leq N}\mathbb{E}\left[\sup_{0\leq t\leq T}\left|X_t^{i,N,M}\right|^p\right]<\infty,$$

then the estimates (9.3.3) *and* (9.3.4) *hold for those p as well.*

Proof of Lemma 9.3.3. Using Hölder's inequality we obtain for any $p\geq 2$

$$\begin{aligned}
&\Big|\int_{\kappa(t)}^t b_M\big(\kappa(s),X_{\kappa(s)}^{i,N,M},\mu_{\kappa(s)}^{X,N,M}\big)\mathrm{d}s\Big|^p\\
&\leq\Big(\int_{\kappa(t)}^t|1|^{\frac{p}{p-1}}\mathrm{d}s\Big)^{p-1}\int_{\kappa(t)}^t\Big|b_M\big(\kappa(s),X_{\kappa(s)}^{i,N,M},\mu_{\kappa(s)}^{X,N,M}\big)\Big|^p\mathrm{d}s\\
&\leq\left(\frac{T}{M}\right)^{p-1}\frac{T}{M}M^{p\alpha}\\
&\leq T^pM^{-p/2},
\end{aligned}\tag{9.3.5}$$

since $|b_M|\leq M^\alpha$ and $\alpha\leq 1/2$. It is easy to see that in the case of $p\in(0,2]$

$$\begin{aligned}
&\mathbb{E}\left[\left|X_t^{i,N,M}-X_{\kappa(t)}^{i,N,M}\right|^p\right]\\
&\leq\mathbb{E}\Big[\Big|\int_{\kappa(t)}^t b_M\big(\kappa(s),X_{\kappa(s)}^{i,N,M},\mu_{\kappa(s)}^{X,N,M}\big)\mathrm{d}s+\int_{\kappa(t)}^t\sigma\big(\kappa(s),X_{\kappa(s)}^{i,N,M},\mu_{\kappa(s)}^{X,N,M}\big)\mathrm{d}W_s^i\Big|^2\Big]^{\frac{p}{2}}\\
&\leq 2^{p/2}\mathbb{E}\Big[\Big|\int_{\kappa(t)}^t b_M\big(\kappa(s),X_{\kappa(s)}^{i,N,M},\mu_{\kappa(s)}^{X,N,M}\big)\mathrm{d}s\Big|^2+\Big|\int_{\kappa(t)}^t\sigma\big(\kappa(s),X_{\kappa(s)}^{i,N,M},\mu_{\kappa(s)}^{X,N,M}\big)\mathrm{d}W_s^i\Big|^2\Big]^{\frac{p}{2}}\\
&\leq 2^{p/2}\mathbb{E}\Big[T^2M^{-1}+\Big|\int_{\kappa(t)}^t\sigma\big(\kappa(s),X_{\kappa(s)}^{i,N,M},\mu_{\kappa(s)}^{X,N,M}\big)\mathrm{d}W_s^i\Big|^2\Big]^{\frac{p}{2}},
\end{aligned}$$

and due to Itô's isometry and Lemma 9.3.2 for C independent of M and i

$$\begin{aligned}
&\mathbb{E}\Big[\Big|\int_{\kappa(t)}^t\sigma\big(\kappa(s),X_{\kappa(s)}^{i,N,M},\mu_{\kappa(s)}^{X,N,M}\big)\mathrm{d}W_s^i\Big|^2\Big]\\
&\leq\mathbb{E}\Big[\int_{\kappa(t)}^t K\Big(1+|X_{\kappa(s)}^{i,N,M}|^2+\frac{1}{N}\sum_{j=1}^N|X_{\kappa(s)}^{j,N,M}|^2\Big)\mathrm{d}s\Big]\\
&\leq\sup_{1\leq i\leq N}\sup_{s\in[\kappa(t),t]}\mathbb{E}\left[\frac{T}{M}K\left(1+|X_s^{i,N,M}|^2+|X_s^{i,N,M}|^2\right)\right]\leq CM^{-1},
\end{aligned}$$

which combined with our previous bound yields

$$\sup_{0\leq t\leq T}\mathbb{E}\left[\left|X_t^{i,N,M}-X_{\kappa(t)}^{i,N,M}\right|^p\right]\leq CM^{-p/2},$$

for all $p\in(0,2]$, hence we obtain (9.3.3). If additionally $\sup_{M\geq 1}\sup_{1\leq i\leq N}\mathbb{E}\big[\sup_{0\leq t\leq T}|X_t^{i,N,M}|^p\big]<$

∞ for some $p > 2$, then

$$\begin{aligned}
&\mathbb{E}\Big[\Big|X_t^{i,N,M} - X_{\kappa(t)}^{i,N,M}\Big|^p\Big] \\
&\leq C\mathbb{E}\Big[\Big|\int_{\kappa(t)}^t b_M\big(\kappa(s), X_{\kappa(s)}^{i,N,M}, \mu_{\kappa(s)}^{X,N,M}\big)\mathrm{d}s\Big|^p + \Big|\int_{\kappa(t)}^t \sigma\big(\kappa(s), X_{\kappa(s)}^{i,N,M}, \mu_{\kappa(s)}^{X,N,M}\big)\mathrm{d}W_s^i\Big|^p\Big] \\
&\leq C\mathbb{E}\Big[T^p M^{p/2} + \Big|\int_{\kappa(t)}^t \sigma\big(\kappa(s), X_{\kappa(s)}^{i,N,M}, \mu_{\kappa(s)}^{X,N,M}\big)^2\mathrm{d}s\Big|^{p/2}\Big],
\end{aligned}$$

by the estimate (9.3.5) and the Burkholder-Davis-Gundy inequality. Since furthermore,

$$\begin{aligned}
&\mathbb{E}\Big[\Big|\int_{\kappa(t)}^t \sigma\big(\kappa(s), X_{\kappa(s)}^{i,N,M}, \mu_{\kappa(s)}^{X,N,M}\big)^2\mathrm{d}s\Big|^{p/2}\Big] \\
&\qquad\leq \mathbb{E}\Big[\Big(\frac{T}{M}\Big)^{p/2} \sup_{s\in[\kappa(t),t]} K\Big(1 + |X_s^{i,N,M}|^p + \Big(\frac{1}{N}\sum_{j=1}^N |X_s^{j,N,M}|^2\Big)^{p/2}\Big)\Big] \\
&\qquad\leq \left(\frac{T}{M}\right)^{p/2} K\left(1 + \mathbb{E}\left[\sup_{0\leq t\leq T}\left|X_t^{i,N,M}\right|^p\right] + \sup_{1\leq j\leq N}\mathbb{E}\left[\sup_{0\leq t\leq T}\left|X_t^{j,N,M}\right|^p\right]\right) \\
&\qquad\leq CM^{-p/2},
\end{aligned}$$

where we have used that $(\frac{1}{N}\sum_{j=1}^N |X_s^{j,N,M}|^2)^{p/2} \leq \frac{1}{N}\sum_{j=1}^N |X_s^{j,N,M}|^p$, since $p > 2$ and that the particles are identically distributed. Hence we get the desired result here as well.

Finally, using the above results and that $\alpha \leq 1/2$, we obtain for any $p \geq 0$ such that $\mathbb{E}[|X_t^{i,N,M} - X_{\kappa(t)}^{i,N,M}|^p] \leq CM^{-p/2}$,

$$\mathbb{E}\left[\left|X_t^{i,N,M} - X_{\kappa(t)}^{i,N,M}\right|^p \left|b_M\big(\kappa(t), X_{\kappa(t)}^{i,N,M}, \mu_{\kappa(t)}^{X,N,M}\big)\right|^p\right] \leq \mathbb{E}\left[\left|X_t^{i,N,M} - X_{\kappa(t)}^{i,N,M}\right|^p\right] M^{p\alpha} \leq C,$$

holds for any $t \in [0, T]$ and $1 \leq i \leq N$, which completes the proof. □

Lemma 9.3.4. *Suppose that Assumption 8.2.1 and 9.1.1 are fulfilled, then for every* $p \geq 2$ *with* $X_0 \in L^p(\mathbb{R}^d)$ *there exists a constant* C *such that*

$$\sup_{M\geq 1}\sup_{1\leq i\leq N}\mathbb{E}\left[\sup_{0\leq t\leq T}\left|X_t^{i,N,M}\right|^p\right] < C.$$

Proof. Define $\hat{p} \geq 2$ such that $\mathbb{E}[|X_0|^{\hat{p}}] < \infty$ and note that if $\hat{p} < 2$ we have nothing to prove by using Lemma 9.3.3.

We use an inductive argument and start with $p = 2$. In every step we set $q = 2p \wedge \hat{p}$. By Itô's formula we have

$$\begin{aligned}
\mathbb{E}\Big[\sup_{0\leq s\leq t}|X_s^{i,N,M}|^q\Big] \leq C\Big(1 &+ \mathbb{E}\Big[|X_0^{i,N,M}|^q\Big] + \int_0^t \mathbb{E}\Big[|X_{\kappa(s)}^{i,N,M}|^q\Big]\mathrm{d}s \\
&+ \int_0^t \mathbb{E}\Big[|X_s^{i,N,M} - X_{\kappa(s)}^{i,N,M}|^{q/2}|b_M\big(\kappa(s), X_{\kappa(s)}^{i,N,M}, \mu_{\kappa(s)}^{X,N,M}\big)|^{q/2}\Big]\mathrm{d}s \\
&+ \mathbb{E}\Big[\sup_{0\leq s\leq t}\Big|\int_0^s X_u^{i,N,M}\sigma\big(\kappa(u), X_{\kappa(u)}^{i,N,M}, \mu_{\kappa(u)}^{X,N,M}\big)\mathrm{d}W_u^i\Big|^{q/2}\Big]\Big),
\end{aligned}$$

and the application of the Burkholder-Davis-Gundy inequality and Lemma 9.3.3 with[*] $q/2$ yields

$$\begin{aligned}
\mathbb{E}\left[\sup_{0\leq s\leq t}|X_s^{i,N,M}|^q\right] \leq &C\Big(1 + \mathbb{E}\Big[\Big|X_0^{i,N,M}\Big|^q\Big] + \int_0^t \mathbb{E}\Big[\sup_{0\leq u\leq s}|X_u^{i,N,M}|^q\Big]\mathrm{d}s \\
&+ \mathbb{E}\Big[\Big(\int_0^t |X_s^{i,N,M}|^2\left|\sigma\big(\kappa(s), X_{\kappa(s)}^{i,N,M}, \mu_{\kappa(s)}^{X,N,M}\big)\right|^2\mathrm{d}s\Big)^{q/4}\Big]\Big),
\end{aligned}$$

[*] Observe that Lemma 9.3.3 holds for the current value of p and since $q = 2p \wedge \hat{p}$ it implies that it holds for $q/2$.

where C denotes in each case a constant that is independent of M. With Young's inequality in the form $ab \leq \frac{1}{2C}a^2 + \frac{C}{2}b^2$, Hölder's inequality and the estimate for σ we obtain

$$
\begin{aligned}
&\mathbb{E}\left[\sup_{0\leq s\leq t}|X_s^{i,N,M}|^q\right] \\
\leq& C\left(1+\mathbb{E}\left[\left|X_0^{i,N,M}\right|^q\right] + \int_0^t \mathbb{E}\left[\sup_{0\leq u\leq s}|X_u^{i,N,M}|^q\right]\mathrm{d}s + \frac{1}{2C}\mathbb{E}\left[\sup_{0\leq s\leq t}|X_s^{i,N,M}|^q\right]\right. \\
&\left.+\frac{C}{2}\mathbb{E}\left[\int_0^t\left|\sigma\left(\kappa(s), X_{\kappa(s)}^{i,N,M}, \mu_{\kappa(s)}^{X,N,M}\right)\right|^q\mathrm{d}s\right]\right) \\
\leq& C\Big(1+\mathbb{E}\left[\left|X_0^{i,N,M}\right|^q\right] + \int_0^t \mathbb{E}\left[\sup_{0\leq u\leq s}|X_u^{i,N,M}|^q\right]\mathrm{d}s + \frac{1}{2C}\mathbb{E}\left[\sup_{0\leq s\leq t}|X_s^{i,N,M}|^q\right] \\
&+\frac{C}{2}\int_0^t\mathbb{E}\Big[\sup_{0\leq u\leq s}K\Big(1+|X_u^{i,N,M}|^q + \Big(\frac{1}{N}\sum_{j=1}^N|X_u^{j,N,M}|^2\Big)^{q/2}\Big)\Big]\mathrm{d}s\Big).
\end{aligned}
$$

Taking the $\frac{1}{2}\mathbb{E}[\sup_{0\leq s\leq t}|X_s^{i,N,M}|^q]$ term to the LHS taking the sup over i on both sides we obtain

$$
\sup_{1\leq i\leq N}\mathbb{E}\left[\sup_{0\leq s\leq t}|X_s^{i,N,M}|^q\right] \leq C\left(1+\mathbb{E}\left[|X_0^{i,N,M}|^q\right] + \int_0^t\sup_{1\leq i\leq N}\mathbb{E}\left[\sup_{0\leq u\leq s}|X_u^{i,N,M}|^q\right]\mathrm{d}s\right) < \infty,
$$

and thus the application of Grönwall's lemma yields that

$$
\sup_{1\leq i\leq N}\mathbb{E}\Big[\sup_{0\leq t\leq T}|X_t^{i,N,M}|^q\Big] < C, \tag{9.3.6}
$$

holds for some positive constant C which dependent on $\mathbb{E}[|X_0^i|^q]$ but is independent of N and M.

Since (9.3.6) is proven for q we can set $p = q$ and use this result in the next step of the iteration. Since the new q is at most twice as much as p, Lemma 9.3.3 can again be applied for $q/2$. This iteration gets repeated until $q = \hat{p}$. □

Now we can complete the proof of Proposition 9.1.4.

Proof of Proposition 9.1.4. Using Itô's formula we observe,

$$
\begin{aligned}
\left|\Delta X_t^{i,N,M}\right|^2 =& \int_0^t 2\langle\Delta X_s^{i,N,M}, \left(b\left(s, X_s^{i,N}, \mu_s^{X,N}\right) - b_M\left(\kappa(s), X_{\kappa(s)}^{i,N,M}, \mu_{\kappa(s)}^{X,N,M}\right)\right)\rangle\mathrm{d}s \\
&+\sum_{a=1}^l\int_0^t|\sigma_a\left(s, X_s^{i,N}, \mu_s^{X,N}\right) - \sigma_a\left(\kappa(s), X_{\kappa(s)}^{i,N,M}, \mu_{\kappa(s)}^{X,N,M}\right)|^2\mathrm{d}s \\
&+\int_0^t 2\langle\Delta X_s^{i,N,M}, \left(\sigma\left(s, X_s^{i,N}, \mu_s^{X,N}\right) - \sigma\left(\kappa(s), X_{\kappa(s)}^{i,N,M}, \mu_{\kappa(s)}^{X,N,M}\right)\right)\mathrm{d}W_s^i\rangle.
\end{aligned}
$$

Furthermore observe that

$$
\begin{aligned}
&\langle X_s^{i,N} - X_s^{i,N,M}, b\big(s, X_s^{i,N}, \mu_s^{X,N}\big) - b_M\big(\kappa(s), X_{\kappa(s)}^{i,N,M}, \mu_{\kappa(s)}^{X,N,M}\big)\rangle \\
=&\langle\Delta X_s^{i,N,M}, b\big(s, X_s^{i,N}, \mu_s^{X,N}\big) - b\big(s, X_s^{i,N,M}, \mu_s^{X,N}\big)\rangle \\
&+\langle\Delta X_s^{i,N,M}, b\big(s, X_s^{i,N,M}, \mu_s^{X,N}\big) - b\big(s, X_s^{i,N,M}, \mu_s^{X,N,M}\big)\rangle \\
&+\langle\Delta X_s^{i,N,M}, b\big(s, X_s^{i,N,M}, \mu_s^{X,N,M}\big) - b\big(\kappa(s), X_s^{i,N,M}, \mu_s^{X,N,M}\big)\rangle \\
&+\langle\Delta X_s^{i,N,M}, b\big(\kappa(s), X_s^{i,N,M}, \mu_s^{X,N,M}\big) - b\big(\kappa(s), X_{\kappa(s)}^{i,N,M}, \mu_s^{X,N,M}\big)\rangle \\
&+\langle\Delta X_s^{i,N,M}, b\big(\kappa(s), X_{\kappa(s)}^{i,N,M}, \mu_s^{X,N,M}\big) - b\big(\kappa(s), X_{\kappa(s)}^{i,N,M}, \mu_{\kappa(s)}^{X,N,M}\big)\rangle \\
&+\langle\Delta X_s^{i,N,M}, b\big(\kappa(s), X_{\kappa(s)}^{i,N,M}, \mu_{\kappa(s)}^{X,N,M}\big) - b_M\big(\kappa(s), X_{\kappa(s)}^{i,N,M}, \mu_{\kappa(s)}^{X,N,M}\big)\rangle,
\end{aligned}
$$

where we estimate every term on the right hand side in the following. Due to Assumption 8.2.1 we have

$$\langle \Delta X_s^{i,N,M}, b\left(s, X_s^{i,N}, \mu_s^{X,N}\right) - b\left(s, X_s^{i,N,M}, \mu_s^{X,N}\right)\rangle \leq L_b \left|\Delta X_s^{i,N,M}\right|^2,$$

and

$$\begin{aligned}&\langle \Delta X_s^{i,N,M}, b\left(s, X_s^{i,N,M}, \mu_s^{X,N}\right) - b\left(s, X_s^{i,N,M}, \mu_s^{X,N,M}\right)\rangle \\ &\quad \leq \left|\Delta X_s^{i,N,M}\right| \left|W^{(2)}\left(\mu_s^{X,N}, \mu_s^{X,N,M}\right)\right| \\ &\quad \leq \left|\Delta X_s^{i,N,M}\right| \frac{1}{\sqrt{N}} \Big(\sum_{j=1}^N |\Delta X_s^{j,N,M}|^2\Big)^{1/2} \\ &\quad \leq \frac{1}{2}\left|\Delta X_s^{i,N,M}\right|^2 + \frac{1}{2}\frac{1}{N}\sum_{j=1}^N |\Delta X_s^{j,N,M}|^2,\end{aligned}$$

and

$$\begin{aligned}&\langle \Delta X_s^{i,N,M}, b\left(s, X_s^{i,N,M}, \mu_s^{X,N,M}\right) - b\left(\kappa(s), X_s^{i,N,M}, \mu_s^{X,N,M}\right)\rangle \\ &\qquad \leq C\left|\Delta X_s^{i,N,M}\right| |s-\kappa(s)|^{1/2} \leq \frac{1}{2}\left|\Delta X_s^{i,N,M}\right|^2 + CM^{-1}.\end{aligned}$$

Further,

$$\begin{aligned}&\langle \Delta X_s^{i,N,M}, b\big(\kappa(s), X_s^{i,N,M}, \mu_s^{X,N,M}\big) - b\big(\kappa(s), X_{\kappa(s)}^{i,N,M}, \mu_s^{X,N,M}\big)\rangle \\ &\leq \frac{1}{2}\left|\Delta X_s^{i,N,M}\right|^2 + \frac{1}{2}\left|b\big(\kappa(s), X_s^{i,N,M}, \mu_s^{X,N,M}\big) - b\big(\kappa(s), X_{\kappa(s)}^{i,N,M}, \mu_s^{X,N,M}\big)\right|^2.\end{aligned}$$

As we will take supremum over time and expected values we can furthermore estimate the change in drift by using the polynomial growth of b with rate q, Hölder's inequality,

$$\begin{aligned}&\mathbb{E}\left[\sup_{u\in[0,t]}\int_0^u \left|b\big(\kappa(s), X_s^{i,N,M}, \mu_s^{X,N,M}\big) - b\big(\kappa(s), X_{\kappa(s)}^{i,N,M}, \mu_s^{X,N,M}\big)\right|^2 \mathrm{d}s\right] \\ &\leq \int_0^t \mathbb{E}\left[L\left(1+\left|X_s^{i,N,M}\right|^q + \left|X_{\kappa(s)}^{i,N,M}\right|^q\right)^2 \left|X_s^{i,N,M} - X_{\kappa(s)}^{i,N,M}\right|^2\right]\mathrm{d}s \\ &\leq \int_0^t \sqrt{\mathbb{E}\left[L\left(1+\left|X_s^{i,N,M}\right|^q + \left|X_{\kappa(s)}^{i,N,M}\right|^q\right)^4\right]\mathbb{E}\left[\left|X_s^{i,N,M} - X_{\kappa(s)}^{i,N,M}\right|^4\right]}\mathrm{d}s.\end{aligned}$$

By Lemma 9.3.4 we obtain,

$$\sup_{M\geq 1}\sup_{1\leq i\leq N}\mathbb{E}\left[\sup_{0\leq t\leq T}\left|X_t^{i,N,M}\right|^{4q}\right] \leq 1 + \sup_{M\geq 1}\sup_{1\leq i\leq N}\mathbb{E}\left[\sup_{0\leq t\leq T}\left|X_t^{i,N,M}\right|^{4(1+q)}\right] < \infty,$$

and using Lemma 9.3.3 to bound the other expectation yields the following,

$$\mathbb{E}\left[\sup_{u\in[0,t]}\int_0^u \left|b\big(\kappa(s), X_s^{i,N,M}, \mu_s^{X,N,M}\big) - b\big(\kappa(s), X_{\kappa(s)}^{i,N,M}, \mu_s^{X,N,M}\big)\right|^2 \mathrm{d}s\right] \leq CM^{-1}.$$

Again Assumption 8.2.1 yields

$$\begin{aligned}
&\langle \Delta X_s^{i,N,M}, b\big(\kappa(s), X_{\kappa(s)}^{i,N,M}, \mu_s^{X,N,M}\big) - b\big(\kappa(s), X_{\kappa(s)}^{i,N,M}, \mu_{\kappa(s)}^{X,N,M}\big)\rangle \\
&\leq \left|\Delta X_s^{i,N,M}\right| \frac{1}{\sqrt{N}} \Big(\sum_{j=1}^{N} |X_s^{j,N,M} - X_{\kappa(s)}^{j,N,M}|^2 \Big)^{1/2} \\
&\leq \frac{1}{2} \left|\Delta X_s^{i,N,M}\right|^2 + \frac{1}{2}\frac{1}{N} \sum_{j=1}^{N} \left| X_s^{j,N,M} - X_{\kappa(s)}^{j,N,M} \right|^2,
\end{aligned}$$

and the definition of b_M that

$$\begin{aligned}
&\langle \Delta X_s^{i,N,M}, b\big(\kappa(s), X_{\kappa(s)}^{i,N,M}, \mu_{\kappa(s)}^{X,N,M}\big) - b_M\big(\kappa(s), X_{\kappa(s)}^{i,N,M}, \mu_{\kappa(s)}^{X,N,M}\big)\rangle \\
&\leq \frac{1}{2} \left|\Delta X_s^{i,N,M}\right|^2 + \frac{1}{2} \left| b\big(\kappa(s), X_{\kappa(s)}^{i,N,M}, \mu_{\kappa(s)}^{X,N,M}\big) - b_M\big(\kappa(s), X_{\kappa(s)}^{i,N,M}, \mu_{\kappa(s)}^{X,N,M}\big) \right|^2 \\
&\leq \frac{1}{2} \left|\Delta X_s^{i,N,M}\right|^2 + \frac{1}{2} M^{-2\alpha} \left| b\left(\kappa(s), X_{\kappa(s)}^{i,N,M}, \mu_{\kappa(s)}^{X,N,M}\right) \right|^4 \\
&\leq \frac{1}{2} \left|\Delta X_s^{i,N,M}\right|^2 + CM^{-2\alpha} \Big(1 + \left| X_{\kappa(s)}^{i,N,M} \right|^{4(1+q)} + \Big(\frac{1}{N} \sum_{j=1}^{N} \left| X_{\kappa(s)}^{j,N,M} \right|^2 \Big)^2 \Big),
\end{aligned}$$

where q is again the polynomial growth rate of b. For the stochastic integral the Burkholder-Davis-Gundy inequality yields

$$\begin{aligned}
&\mathbb{E}\left[\sup_{u\in[0,t]} \int_0^u 2\langle \Delta X_s^{i,N,M}, \left(\sigma\left(s, X_s^{i,N}, \mu_s^{X,N}\right) - \sigma\left(\kappa(s), X_{\kappa(s)}^{i,N,M}, \mu_{\kappa(s)}^{X,N,M}\right)\right) \mathrm{d}W_s^i \rangle \right] \\
&\leq \mathbb{E}\left[\Big(C \int_0^t \Big(\sum_{a=1}^{l} |\sigma_a(s, X_s^{i,N}, \mu_s^{X,N}) - \sigma_a(s, X_{\kappa(s)}^{i,N,M}, \mu_{\kappa(s)}^{X,N,M})|^2 \Big) |\Delta X_s^{i,N,M}|^2 \mathrm{d}s \Big)^{\frac{1}{2}} \right] \\
&\leq \mathbb{E}\Big[\frac{1}{2} \sup_{u\in[0,t]} |\Delta X_u^{i,N,M}|^2 + C \int_0^t \sum_{a=1}^{l} \left| \sigma_a(s, X_s^{i,N}, \mu_s^{X,N}) - \sigma_a(\kappa(s), X_{\kappa(s)}^{i,N,M}, \mu_{\kappa(s)}^{X,N,M}) \right|^2 \mathrm{d}s \Big].
\end{aligned}$$

and

$$\begin{aligned}
&\left| \sigma_a\left(s, X_s^{i,N}, \mu_s^{X,N}\right) - \sigma_a\left(\kappa(s), X_{\kappa(s)}^{i,N,M}, \mu_{\kappa(s)}^{X,N,M}\right) \right|^2 \\
&\leq C\left|s - \kappa(s)\right| + C\left| X_s^{i,N} - X_{\kappa(s)}^{i,N,M} \right|^2 + CW^{(2)}\left(\mu_s^{X,N}, \mu_{\kappa(s)}^{X,N,M}\right)^2 \\
&\leq CM^{-1} + C\left| X_s^{i,N} - X_{\kappa(s)}^{i,N,M} \right|^2 + \frac{C}{N} \sum_{j=1}^{N} \left| X_s^{j,N} - X_{\kappa(s)}^{j,N,M} \right|^2 \\
&\leq CM^{-1} + C\left| X_s^{i,N} - X_{\kappa(s)}^{i,N,M} \right|^2 + \frac{C}{N} \sum_{j=1}^{N} \left(\left|\Delta X_s^{j,N,M}\right|^2 + \left| X_s^{j,N,M} - X_{\kappa(s)}^{j,N,M} \right|^2 \right).
\end{aligned}$$

By putting this together we obtain

$$
\begin{aligned}
&\mathbb{E}\left[\sup_{0\le u\le t}\left|\Delta X_u^{i,N,M}\right|^2\right]\\
&\le C\mathbb{E}\Bigg[\int_0^t \left|\Delta X_s^{i,N,M}\right|^2+\frac{1}{N}\sum_{j=1}^N\left|X_s^{j,N,M}-X_{\kappa(s)}^{j,N,M}\right|^2+M^{-1}+\frac{1}{N}\sum_{j=1}^N\left|\Delta X_s^{j,N,M}\right|^2\\
&+\left|X_s^{i,N,M}-X_{\kappa(s)}^{i,N,M}\right|^2+\left|b\big(\kappa(s),X_s^{i,N,M},\mu_s^{X,N,M}\big)-b\big(\kappa(s),X_{\kappa(s)}^{i,N,M},\mu_s^{X,N,M}\big)\right|^2\\
&+M^{-2\alpha}\Big(1+\left|X_{\kappa(s)}^{i,N,M}\right|^{4(1+q)}\Big)+M^{-2\alpha}\Big(\frac{1}{N}\sum_{j=1}^N\left|X_{\kappa(s)}^{j,N,M}\right|^2\Big)^2\mathrm{d}s\\
&+\int_0^t\langle\Delta X_s^{i,N,M},\Big(\sigma\left(s,X_s^{i,N},\mu_s^{X,N}\right)-\sigma\left(\kappa(s),X_{\kappa(s)}^{i,N,M},\mu_{\kappa(s)}^{X,N,M}\right)\Big)\,\mathrm{d}W_s^i\rangle\Bigg]\\
&\le C\Big(\int_0^t\mathbb{E}\Big[\sup_{0\le u\le s}\left|\Delta X_u^{i,N,M}\right|^2\Big]\mathrm{d}s+M^{-2\alpha}+M^{-1}\Big),
\end{aligned}
$$

by Lemma 9.3.4 and since $X^{i,N}$ are identically distributed and $X^{i,N,M}$ are identically distributed for all $i\in\{1,\dots,N\}$. This estimate holds for every i hence we can insert $\sup_{1\le i\le N}$ on both sides giving

$$
\sup_{1\le i\le N}\mathbb{E}\left[\sup_{0\le u\le t}\left|\Delta X_u^{i,N,M}\right|^2\right]\le C\Big(\int_0^t\sup_{1\le i\le N}\mathbb{E}\Big[\sup_{0\le u\le s}\left|\Delta X_u^{i,N,M}\right|^2\Big]\mathrm{d}s+M^{-2\alpha}+M^{-1}\Big)<\infty,
$$

and finally by Grönwall's lemma (using that $\alpha=1/2$),

$$
\sup_{1\le i\le N}\mathbb{E}\left[\sup_{0\le u\le t}\left|X_u^{i,N}-X_u^{i,N,M}\right|^2\right]\le CM^{-1}.
$$

□

9.3.3 Proof of Implicit Convergence

The main goal here is to prove Proposition 9.1.8. We loosely follow [MS13], however, due to the extra dependencies on time and measure and further allowing for random initial conditions we require more refined arguments. We take N as some fixed positive integer. Before considering the implicit scheme, let us show a result on the particle system (8.2.2).

Proposition 9.3.5. *Let Assumption 8.2.1, 9.1.1 and H1 (in Assumption 9.1.6) hold, further, let $X_0\in L^2(\mathbb{R}^d)$. Then the following bounds hold,*

$$
\sup_{1\le i\le N}\mathbb{E}[|X_T^{i,N}|^2]\le\big(\mathbb{E}[|X_0|^2]+2\alpha T\big)\exp(2\beta T),
$$

and for $\tau_m^i=\inf\{t\ge0:\ |X_t^{i,N}|>m\}$

$$
\sup_{1\le i\le N}\mathbb{P}(\tau_m^i\le T)\le\frac{1}{m^2}\big(\mathbb{E}[|X_0|^2]+2\alpha T\big)\exp(2\beta T)\,.
$$

Proof. Firstly, let us consider the stopped process $X_{T\wedge\tau_m^i}^{i,N}$. Applying Itô to the square of this process and taking expectations yields

$$
\begin{aligned}
\mathbb{E}[|X_{T\wedge\tau_m^i}^{i,N}|^2]&=\mathbb{E}[|X_0^i|^2]+\mathbb{E}\Big[\int_0^{T\wedge\tau_m^i}2\langle X_s^{i,N},b(s,X_s^{i,N},\mu_s^{X,N})\rangle+\sum_{a=1}^l|\sigma_a(s,X_s^{i,N},\mu_s^{X,N})|^2\mathrm{d}s\Big]\\
&\le\mathbb{E}[|X_0^i|^2]+2\alpha T+\int_0^T2\beta\mathbb{E}[X_{s\wedge\tau_m^i}^{i,N}]\mathrm{d}s\le\big(\mathbb{E}[|X_0^i|^2]+2\alpha T\big)e^{2\beta T},
\end{aligned}
$$

where we have used the growth and stopping condition to remove the martingale term, then the monotone growth, uniform boundedness of b in the measure component b and Grönwall's inequality to obtain the result.

Noting that the following lower bound also holds,

$$\mathbb{E}[|X^{i,N}_{T\wedge\tau^i_m}|^2] \geq m^2 \mathbb{P}(\tau^i_m \leq T)\,,$$

hence we obtain,

$$\mathbb{P}(\tau^i_m \leq T) \leq \frac{1}{m^2}\big(\mathbb{E}[|X^i_0|^2] + 2\alpha T\big)\exp(2\beta T)\,.$$

Further, since $\lim_{m\to\infty}|X^{i,N}_{T\wedge\tau^i_m}| = |X^{i,N}_T|$, we obtain by Fatou's lemma,

$$\mathbb{E}[|X^{i,N}_T|^2] \leq \liminf_{m\to\infty}\mathbb{E}[|X^{i,N}_{T\wedge\tau^i_m}|^2] \leq \big(\mathbb{E}[|X^i_0|^2] + 2\alpha T\big)\exp(2\beta T)\,.$$

The result then follows by noting that $\mathbb{E}[|X^i_0|^2] = \mathbb{E}[|X_0|^2]$ and hence the bounds are independent of i, so we obtain the result for the supremum over i. □

Let us now return to the implicit scheme. At each time step t_i and for each particle i one needs to solve a fixed point equation,

$$\tilde{X}^{i,N,M}_{t_{k+1}} - b\Big(t_k, \tilde{X}^{i,N,M}_{t_{k+1}}, \tilde{\mu}^{X,N,M}_{t_k}\Big)h = \tilde{X}^{i,N,M}_{t_k} + \sigma\Big(t_k, \tilde{X}^{i,N,M}_{t_k}, \tilde{\mu}^{X,N,M}_{t_k}\Big)\Delta W^i_{t_k},$$

this leads us to consider a function F

$$F(t,x,\mu) := x - b(t,x,\mu)h. \tag{9.3.7}$$

For the implicit scheme to have a solution the function F must have a unique inverse. The following lemma is crucial in proving convergence of the implicit scheme.

Lemma 9.3.6. *Let Assumption 8.2.1, 9.1.1 and H1 (in Assumption 9.1.6) hold and fix $h^* < 1/\max(L_b, 2\beta)$. Further, let $0 < h \leq h^*$ and take any $t \in [0,T]$ and $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ fixed, then for all $y \in \mathbb{R}^d$, there exists a unique x such that $F(t,x,\mu) = y$. Hence the fixed point problem in* (9.0.2) *is well defined.*

Moreover, for all $t \in [0,T]$ and $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ the following bound holds,

$$|x|^2 \leq (1-2h\beta)^{-1}(|F(t,x,\mu)|^2 + 2h\alpha)\,,$$

and for any $i \geq 1$ the following recursive bound holds,

$$\begin{aligned}
&|F(t_k, \tilde{X}^{i,N,M}_{t_{k+1}}, \tilde{\mu}^{X,N,M}_{t_k})|^2\\
&\leq |F(t_{k-1}, \tilde{X}^{i,N,M}_{t_k}, \tilde{\mu}^{X,N,M}_{t_{k-1}})|^2 + \Big(\sum_{a=1}^{l}|\sigma_a(t_k, \tilde{X}^{i,N,M}_{t_k}, \tilde{\mu}^{X,N,M}_{t_k})||(\Delta W^i_{t_k})_a|\Big)^2\\
&\quad + 2h\alpha + 2h\beta|\tilde{X}^{i,N,M}_{t_k}|^2 + 2\langle \tilde{X}^{i,N,M}_{t_k}, \sigma(t_k, \tilde{X}^{i,N,M}_{t_k}, \tilde{\mu}^{X,N,M}_{t_k})\Delta W^i_{t_k}\rangle\,,
\end{aligned} \tag{9.3.8}$$

where $\big(\Delta W^i_{t_k}\big)_a$ is the ath entry of the vector.

Proof. Let us first prove there exists a unique solution to (9.3.7), in the sense that for all $t \in [0,T]$ and $\mu \in \mathcal{P}_2(\mathbb{R}^d)$ fixed, then there exists a unique $x \in \mathbb{R}^d$ such that $F(t,x,\mu) = y$ for a given $y \in \mathbb{R}^d$, provided $0 < h < h^*$. This is a classical problem considered in [Zei90, p.557] or see [LdRS15, p.2596], which requires F to be continuous, monotone and coercive (in x). Clearly, since b is continuous, one has F is continuous. For monotonicity in F,

$$\begin{aligned}
\langle x - x', F(t,x,\mu) - F(t,x',\mu)\rangle &= |x-x'|^2 - \langle x-x', b(t,x,\mu)h - b(t,x',\mu)h\rangle\\
&\geq |x-x'|^2(1-L_b h)\,,
\end{aligned}$$

which is clearly > 0 provided $h < 1/L_b$. Coercivity follows similarly by the monotone growth condition in b,

$$\langle x, F(t,x,\mu)\rangle \geq |x|^2 - h(\alpha + \beta|x|^2)\,,$$

therefore,

$$\lim_{|x|\to\infty} \frac{\langle x, F(t,x,\mu)\rangle}{|x|} = \infty, \quad \text{for } h < 1/\beta.$$

Hence $F(t,x,\mu) = y$ has a unique solution for F defined in (9.3.7) and therefore the numerical scheme (9.0.2) is well defined.

To show x is bounded by $F(\cdot,x,\cdot)$, again fix some $t \in [0,T]$ and $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, then,

$$\begin{aligned}|F(t,x,\mu)|^2 &= |x|^2 - 2\langle x, b(t,x,\mu)\rangle h + |b(t,x,\mu)|^2 h^2 \\ &\geq |x|^2 - 2\langle x, b(t,x,\mu)\rangle h \geq (1-2h\beta)|x|^2 - 2h\alpha.\end{aligned}$$

Since $h < 1/(2\beta)$, we obtain,

$$|x|^2 \leq (1-2h\beta)^{-1}(|F(t,x,\mu)|^2 + 2h\alpha)\,.$$

This result is also useful since it holds for all $t \in [0,T]$ and $\mu \in \mathcal{P}_2(\mathbb{R}^d)$. For the recursive bound it is useful to note,

$$\begin{aligned}F(t_k, \tilde{X}^{i,N,M}_{t_{k+1}}, \tilde{\mu}^{X,N,M}_{t_k}) &= \tilde{X}^{i,N,M}_{t_{k+1}} - b(t_k, \tilde{X}^{i,N,M}_{t_{k+1}}, \tilde{\mu}^{X,N,M}_{t_k})h \\ &= \tilde{X}^{i,N,M}_{t_k} + \sigma(t_k, \tilde{X}^{i,N,M}_{t_k}, \tilde{\mu}^{X,N,M}_{t_k})\Delta W^i_{t_k} \\ &= F(t_{k-1}, \tilde{X}^{i,N,M}_{t_k}, \tilde{\mu}^{X,N,M}_{t_{k-1}}) + b(t_{k-1}, \tilde{X}^{i,N,M}_{t_k}, \tilde{\mu}^{X,N,M}_{t_{k-1}})h \\ &\quad + \sigma(t_k, \tilde{X}^{i,N,M}_{t_k}, \tilde{\mu}^{X,N,M}_{t_k})\Delta W^i_{t_k}.\end{aligned} \tag{9.3.9}$$

Of course, this recursion is only valid for $k \geq 1$, due to the appearance of t_{k-1}. Using this relation observe the following,

$$\begin{aligned}&|F(t_k, \tilde{X}^{i,N,M}_{t_{k+1}}, \tilde{\mu}^{X,N,M}_{t_k})|^2 \\ &\quad = |F(t_{k-1}, \tilde{X}^{i,N,M}_{t_k}, \tilde{\mu}^{X,N,M}_{t_{k-1}})|^2 + |b(t_{k-1}, \tilde{X}^{i,N,M}_{t_k}, \tilde{\mu}^{X,N,M}_{t_{k-1}})|^2 h^2 \\ &\quad + |\sigma(t_k, \tilde{X}^{i,N,M}_{t_k}, \tilde{\mu}^{X,N,M}_{t_k})\Delta W^i_{t_k}|^2 \\ &\quad + 2\langle F(t_{k-1}, \tilde{X}^{i,N,M}_{t_k}, \tilde{\mu}^{X,N,M}_{t_{k-1}}), b(t_{k-1}, \tilde{X}^{i,N,M}_{t_k}, \tilde{\mu}^{X,N,M}_{t_{k-1}})\rangle h \\ &\quad + 2\langle F(t_{k-1}, \tilde{X}^{i,N,M}_{t_k}, \tilde{\mu}^{X,N,M}_{t_{k-1}}) + b(t_{k-1}, \tilde{X}^{i,N,M}_{t_k}, \tilde{\mu}^{X,N,M}_{t_{k-1}})h, \sigma(t_k, \tilde{X}^{i,N,M}_{t_k}, \tilde{\mu}^{X,N,M}_{t_k})\Delta W^i_{t_k}\rangle.\end{aligned}$$

We now look to bound these various terms, by definition of F,

$$\begin{aligned}&2\langle F(t_{k-1}, \tilde{X}^{i,N,M}_{t_k}, \tilde{\mu}^{X,N,M}_{t_{k-1}}), b(t_{k-1}, \tilde{X}^{i,N,M}_{t_k}, \tilde{\mu}^{X,N,M}_{t_{k-1}})\rangle h \\ &\qquad + |b(t_{k-1}, \tilde{X}^{i,N,M}_{t_k}, \tilde{\mu}^{X,N,M}_{t_{k-1}})|^2 h^2 \leq 2\langle \tilde{X}^{i,N,M}_{t_k}, b(t_{k-1}, \tilde{X}^{i,N,M}_{t_k}, \tilde{\mu}^{X,N,M}_{t_{k-1}})\rangle h \\ &\qquad \leq 2h\alpha + 2h\beta|\tilde{X}^{i,N,M}_{t_k}|^2.\end{aligned}$$

Similarly,

$$\begin{aligned}&2\langle F(t_{k-1}, \tilde{X}^{i,N,M}_{t_k}, \tilde{\mu}^{X,N,M}_{t_{k-1}}) + b(t_{k-1}, \tilde{X}^{i,N,M}_{t_k}, \tilde{\mu}^{X,N,M}_{t_{k-1}})h, \sigma(t_k, \tilde{X}^{i,N,M}_{t_k}, \tilde{\mu}^{X,N,M}_{t_k})\Delta W^i_{t_k}\rangle \\ &\qquad = 2\langle \tilde{X}^{i,N,M}_{t_k}, \sigma(t_k, \tilde{X}^{i,N,M}_{t_k}, \tilde{\mu}^{X,N,M}_{t_k})\Delta W^i_{t_k}\rangle.\end{aligned}$$

In order to obtain the desired form we note the following,

$$\sigma(t,x,\mu)\Delta W_t = \sum_{a=1}^{l} \sigma_a(t,x,\mu)(\Delta W_t)_a\,,$$

crucially one observes $(\Delta W_t)_a$ is a scalar, then standard properties of norms yield,

$$|\sigma(t_k, \tilde{X}^{i,N,M}_{t_k}, \tilde{\mu}^{X,N,M}_{t_k})\Delta W^i_{t_k}| \le \sum_{a=1}^{l} |\sigma_a(t_k, \tilde{X}^{i,N,M}_{t_k}, \tilde{\mu}^{X,N,M}_{t_k})||(\Delta W^i_{t_k})_a| \,.$$

The bound on F then follows immediately from these results. □

Let us now show the first moment bound result, as is standard with implicit schemes we firstly do this under a stopping time, hence define,

$$\lambda^i_m = \inf\{k : \, |\tilde{X}^{i,N,M}_{t_k}| > m\}. \tag{9.3.10}$$

One should note that this stopping time does not actually bound $\tilde{X}$ at that point, the best one can do is bound the previous point i.e. for $\lambda^i_m > 0$, we have $|\tilde{X}^{i,N,M}_{\lambda^i_m - 1}| \le m$.

Lemma 9.3.7. *Let Assumption 8.2.1, 9.1.1 and H1 (in Assumption 9.1.6) hold and fix $h^* < 1/\max(L_b, 2\beta)$. Then for any $p \ge 2$ such that $\mathbb{E}[|X_0|^p] = C(p) < \infty$, we also have,*

$$\sup_{1\le i \le N} \mathbb{E}\big[|\tilde{X}^{i,N,M}_{t_k}|^p \mathbb{1}_{\{k \le \lambda^i_m\}}\big] \le C(p,m) \quad \forall k \le M \ \textit{ and } 0 < h \le h^*.$$

Using standard notation, $C(a)$ denotes a constant that can depend on variable a.

Proof. As it turns out the function F in (9.3.7) gives us a useful bound, from (9.3.9) we obtain,

$$|F(t_k, \tilde{X}^{i,N,M}_{t_{k+1}}, \tilde{\mu}^{X,N,M}_{t_k})|^p \le 2^{p-1}\big(|\tilde{X}^{i,N,M}_{t_k}|^p + |\sigma(t_k, \tilde{X}^{i,N,M}_{t_k}, \tilde{\mu}^{X,N,M}_{t_k})\Delta W^i_{t_k}|^p\big).$$

Hence, multiplying with the indicator and taking expected values yields,

$$\begin{aligned}&\mathbb{E}[|F(t_k, \tilde{X}^{i,N,M}_{t_{k+1}}, \tilde{\mu}^{X,N,M}_{t_k})|^p \mathbb{1}_{\{k+1 \le \lambda^i_m\}}] \\ &\qquad \le C(p)\Big(m^p + \mathbb{E}\big[|\sigma(t_k, \tilde{X}^{i,N,M}_{t_k}, \tilde{\mu}^{X,N,M}_{t_k})\Delta W^i_{t_k}|^p \mathbb{1}_{\{k+1 \le \lambda^i_m\}}\big]\Big)\,.\end{aligned}$$

Then using,

$$\begin{aligned}&\mathbb{E}\big[|\sigma(t_k, \tilde{X}^{i,N,M}_{t_k}, \tilde{\mu}^{X,N,M}_{t_k})\Delta W^i_{t_k}|^p \mathbb{1}_{\{k+1 \le \lambda^i_m\}}\big] \\ &\qquad \le \sum_{a=1}^{l} \mathbb{E}[|\sigma_a(t_k, \tilde{X}^{i,N,M}_{t_k}, \tilde{\mu}^{X,N,M}_{t_k})|^{2p} \mathbb{1}_{\{k+1 \le \lambda^i_m\}}] + \mathbb{E}[|(\Delta W^i_{t_k})_a|^{2p}]\,.\end{aligned}$$

Using the bounds on each coefficient of σ, it is straightforward to observe,

$$|\sigma_a(t_k, \tilde{X}^{i,N,M}_{t_k}, \tilde{\mu}^{X,N,M}_{t_k})|^{2p} \le C(p)\big(1 + |\tilde{X}^{i,N,M}_{t_k}|^{2p}\big)\,.$$

Using this bound we obtain,

$$\mathbb{E}[|F(t_k, \tilde{X}^{i,N,M}_{t_{k+1}}, \tilde{\mu}^{X,N,M}_{t_k})|^p \mathbb{1}_{\{k+1 \le \lambda^i_m\}}] \le C(p,m)\,.$$

Rewriting the quantity we wish to bound as

$$\mathbb{E}\big[|\tilde{X}^{i,N,M}_{t_k}|^p \mathbb{1}_{\{k \le \lambda^i_m\}}\big] = \mathbb{E}\big[|\tilde{X}^{i,N,M}_{t_{k+1}}|^p \mathbb{1}_{\{k+1 \le \lambda^i_m\}}\big] + \mathbb{E}\big[|\tilde{X}^{i,N,M}_{t_0}|^p \mathbb{1}_{\{k=0,\ \lambda^i_m=0\}}\big] \le C(p,m)\,,$$

where the inequality follows from Lemma 9.3.6, our bound on F, and the assumption that $X_0 \in L^p(\mathbb{R}^d)$. Again, the corresponding bound is independent of the choice of i, hence the result holds for the supremum over i. □

Although the previous bound is useful, the presence of the stopping time is inconvenient, we therefore remove it and show the second moment is bounded.

Proposition 9.3.8. *Let Assumption 8.2.1, 9.1.1 and H1 (in Assumption 9.1.6) hold and fix $h^* < 1/\max(L_b, 2\beta)$. Further assume that $X_0 \in L^4(\mathbb{R}^d)$. Then,*

$$\sup_{1\le i\le N}\ \sup_{0<h\le h^*}\ \sup_{0\le k\le M} \mathbb{E}[|\tilde{X}_{t_k}^{i,N,M}|^2] \le C\,.$$

Proof. Firstly let us take a nonnegative integer K, such that $Kh \le T$. Now let us consider (9.3.8), one can note that this bound still holds where the F terms are multiplied by $\mathbb{1}_{\{\lambda_m^i>0\}}$ (since both sides are nonnegative and the indicator is bounded above by one). Summing both sides from $k=1$ to $K\wedge\lambda_m^i$, noting that F terms cancel, we obtain,

$$\begin{aligned}
&|F(t_{K\wedge\lambda_m^i}, \tilde{X}_{t_{(K\wedge\lambda_m^i)+1}}^{i,N,M}, \tilde{\mu}_{t_{K\wedge\lambda_m^i}}^{X,N,M})|^2\mathbb{1}_{\{\lambda_m^i>0\}} \\
&\le |F(t_0, \tilde{X}_{t_1}^{i,N,M}, \tilde{\mu}_{t_0}^{X,N,M})|^2\mathbb{1}_{\{\lambda_m^i>0\}} + \sum_{k=1}^{K\wedge\lambda_m^i}\big(2h\alpha + 2h\beta|\tilde{X}_{t_k}^{i,N,M}|^2\mathbb{1}_{\{\lambda_m^i>0\}}\big) \\
&+ \sum_{k=1}^{K\wedge\lambda_m^i}\Big(\sum_{a=1}^{l}|\sigma_a(t_k, \tilde{X}_{t_k}^{i,N,M}, \tilde{\mu}_{t_k}^{X,N,M})||(\Delta W_{t_k}^i)_a|\Big)^2\mathbb{1}_{\{\lambda_m^i>0\}} \\
&+ \sum_{k=1}^{K\wedge\lambda_m^i} 2\langle \tilde{X}_{t_k}^{i,N,M}, \sigma(t_k, \tilde{X}_{t_k}^{i,N,M}, \tilde{\mu}_{t_k}^{X,N,M})\Delta W_{t_k}^i\rangle\mathbb{1}_{\{\lambda_m^i>0\}}\,,
\end{aligned}$$

where we use the convention $\sum_{k=1}^{0}\cdot = 0$. Although the stopping time is useful it is not ideal that it appears on the sum, however, for nonnegative terms it is straightforward to take the stopping time into the coefficients and for the stochastic term we can rewrite as,

$$\begin{aligned}
&\sum_{k=1}^{K\wedge\lambda_m^i} 2\langle \tilde{X}_{t_k}^{i,N,M}, \sigma(t_k, \tilde{X}_{t_k}^{i,N,M}, \tilde{\mu}_{t_k}^{X,N,M})\Delta W_{t_k}^i\rangle \\
&\qquad = \sum_{k=1}^{K} 2\langle \tilde{X}_{t_k}^{i,N,M}, \sigma(t_k, \tilde{X}_{t_k}^{i,N,M}, \tilde{\mu}_{t_k}^{X,N,M})\Delta W_{t_k}^i\rangle\mathbb{1}_{\{k\le\lambda_m^i\}}.
\end{aligned}$$

Taking expectations of this and noting by Lemma 9.3.7 that $\tilde{X}_{t_k}^{i,N,M}\mathbb{1}_{\{k\le\lambda_m^i\}} \in L_{t_k}^4(\mathbb{R}^d)$, hence, this term is a martingale. We therefore obtain the following bound,

$$\begin{aligned}
&\mathbb{E}[|F(t_{K\wedge\lambda_m^i}, \tilde{X}_{t_{(K\wedge\lambda_m^i)+1}}^{i,N,M}, \tilde{\mu}_{t_{K\wedge\lambda_m^i}}^{X,N,M})|^2\mathbb{1}_{\{\lambda_m^i>0\}}] \\
&\le \mathbb{E}\big[|F(t_0, \tilde{X}_{t_1}^{i,N,M}, \tilde{\mu}_{t_0}^{X,N,M})|^2\big] + 2\alpha T + \sum_{k=1}^{K} 2h\beta\mathbb{E}\big[|\tilde{X}_{t_{k\wedge\lambda_m^i}}^{i,N,M}|^2\mathbb{1}_{\{\lambda_m^i>0\}}\big] \\
&+ \sum_{k=1}^{K}\mathbb{E}\Big[\Big(\sum_{a=1}^{l}|\sigma_a(t_{k\wedge\lambda_m^i}, \tilde{X}_{t_{k\wedge\lambda_m^i}}^{i,N,M}, \tilde{\mu}_{t_{k\wedge\lambda_m^i}}^{X,N,M})||(\Delta W_{t_{k\wedge\lambda_m^i}}^i)_a|\Big)^2\mathbb{1}_{\{\lambda_m^i>0\}}\Big]\,.
\end{aligned}$$

The idea is to apply the discrete version of Grönwall's inequality to this (see for example [MPF12, pg 436] or [MS13, Lemma 3.4]), which requires our bound to be in terms of F. Using arguments similar to previously,

$$\begin{aligned}
&\mathbb{E}\Big[\Big(\sum_{a=1}^{l}|\sigma_a(t_{k\wedge\lambda_m^i}, \tilde{X}_{t_{k\wedge\lambda_m^i}}^{i,N,M}, \tilde{\mu}_{t_{k\wedge\lambda_m^i}}^{X,N,M})||(\Delta W_{t_{k\wedge\lambda_m^i}}^i)_a|\Big)^2\mathbb{1}_{\{\lambda_m^i>0\}}\Big] \\
&\le C\sum_{a=1}^{l}\mathbb{E}\Big[|\sigma_a(t_{k\wedge\lambda_m^i}, \tilde{X}_{t_{k\wedge\lambda_m^i}}^{i,N,M}, \tilde{\mu}_{t_{k\wedge\lambda_m^i}}^{X,N,M})|^2|(\Delta W_{t_{k\wedge\lambda_m^i}}^i)_a|^2\mathbb{1}_{\{\lambda_m^i>0\}}\Big] \\
&\le C\sum_{a=1}^{l} h\big(1 + \mathbb{E}\big[|\tilde{X}_{t_{k\wedge\lambda_m^i}}^{i,N,M}|^2\mathbb{1}_{\{\lambda_m^i>0\}}\big]\big)\,,
\end{aligned}$$

where we have used independence of $\sigma(\cdot)\mathbb{1}_{\{\lambda_m^i>0\}}$ and ΔW along with the growth bounds on σ to obtain the final inequality. Combing this with our previous bounds and appealing again to Lemma 9.3.6 (to bound $\tilde{X}$ by F) we obtain,

$$\begin{aligned}
&\mathbb{E}[|F(t_{K\wedge\lambda_m^i}, \tilde{X}^{i,N,M}_{t_{(K\wedge\lambda_m^i)+1}}, \tilde{\mu}^{X,N,M}_{t_{K\wedge\lambda_m^i}})|^2\mathbb{1}_{\{\lambda_m^i>0\}}] \\
&\leq \mathbb{E}\big[|F(t_0, \tilde{X}^{i,N,M}_{t_1}, \tilde{\mu}^{X,N,M}_{t_0})|^2\big] + C + \sum_{k=1}^{K} Ch\mathbb{E}\big[|\tilde{X}^{i,N,M}_{t_{k\wedge\lambda_m^i}}|^2\mathbb{1}_{\{\lambda_m^i>0\}}\big] \\
&\leq \mathbb{E}\big[|F(t_0, \tilde{X}^{i,N,M}_{t_1}, \tilde{\mu}^{X,N,M}_{t_0})|^2\big] + C\Big(1+\frac{h}{1-2h\beta}\Big) \\
&\qquad + \sum_{k=1}^{K} C\frac{h}{1-2h\beta}\mathbb{E}\big[|F(t_{(k\wedge\lambda_m^i)-1}, \tilde{X}^{i,N,M}_{t_{k\wedge\lambda_m^i}}, \tilde{\mu}^{X,N,M}_{t_{(k\wedge\lambda_m^i)-1}})|^2\mathbb{1}_{\{\lambda_m^i>0\}}\big] .
\end{aligned}$$

Applying a discrete version of Grönwall inequality and noting $\sum_{k=1}^K 1 \leq T/h$ yields

$$\begin{aligned}
&\mathbb{E}[|F(t_{K\wedge\lambda_m^i}, \tilde{X}^{i,N,M}_{t_{(K\wedge\lambda_m^i)+1}}, \tilde{\mu}^{X,N,M}_{t_{K\wedge\lambda_m^i}})|^2\mathbb{1}_{\{\lambda_m^i>0\}}] \\
&\leq \Big(\mathbb{E}\big[|F(t_0, \tilde{X}^{i,N,M}_{t_1}, \tilde{\mu}^{X,N,M}_{t_0})|^2\big] + C\Big(1+\frac{h}{1-2h\beta}\Big)\Big)\exp\Big(\frac{C}{1-2h\beta}\Big) .
\end{aligned}$$

Recalling (9.3.9), we can apply the same arguments as previous to obtain the bound

$$\begin{aligned}
\mathbb{E}\big[|F(t_0, \tilde{X}^{i,N,M}_{t_1}, \tilde{\mu}^{X,N,M}_{t_0})|^2\big] &\leq C\mathbb{E}\big[|\tilde{X}^{i,N,M}_{t_0}|^2 + |\sigma(t_0, \tilde{X}^{i,N,M}_{t_0}, \tilde{\mu}^{X,N,M}_{t_0})\Delta W^i_{t_0}|^2\big] \\
&\leq C(1+(1+h)\mathbb{E}[|\tilde{X}^{i,N,M}_{t_0}|^2]) .
\end{aligned}$$

Noting that our bound for F is now independent of m, we can use Fatou's lemma to take the limit and obtain (for $K \geq 1$),

$$\mathbb{E}[|F(t_K, \tilde{X}^{i,N,M}_{t_{K+1}}, \tilde{\mu}^{X,N,M}_{t_K})|^2] \leq C\Big(1+(1+h)\mathbb{E}[|\tilde{X}^{i,N,M}_{t_0}|^2] + \frac{h}{1-2h\beta}\Big)\exp\Big(\frac{C}{1-2h\beta}\Big) .$$

Again by Lemma 9.3.6, the LHS bounds $\tilde{X}^{i,N,M}_{t_{K+1}}$ (with some constant) hence we obtain a bound for $\tilde{X}^{i,N,M}_{t_k}$ for $k \geq 2$. Clearly $\tilde{X}^{i,N,M}_{t_0}$ has second moment (by assumption), therefore we need to obtain a bound for $\tilde{X}^{i,N,M}_{t_1}$. This is not difficult to obtain by again using that we can bound $\tilde{X}$ as follows,

$$\mathbb{E}\big[|\tilde{X}^{i,N,M}_{t_1}|^2\big] \leq \big(1-2h\beta\big)^{-1}\Big(2h\alpha + \mathbb{E}\big[|F(t_0, \tilde{X}^{i,N,M}_{t_1}, \tilde{\mu}^{X,N,M}_{t_0})|^2\big]\Big),$$

then we can apply the same bound on F as above.

In order to complete the proof, we need to also show this bound exists for all i and $0 < h \leq h^*$. One can see immediately that all bounds decrease as h decreases, hence the supremum value is to set $h = h^*$, which is also finite since $h^* < 1/(2\beta)$. The supremum over i follows from the fact that all bounds are independent of i. □

Now that we have established a bound on the second moment, we look to show convergence of this scheme to the true particle system solution. As always with discrete schemes it is beneficial to introduce their continuous counterpart. As it turns out doing it naively for implicit schemes leads to measurability problems, hence one introduces the so-called forward backward scheme

$$\hat{X}^{i,N,M}_{t_{k+1}} = \hat{X}^{i,N,M}_{t_k} + b\left(t_{k-1\vee 0}, \tilde{X}^{i,N,M}_{t_k}, \tilde{\mu}^{X,N,M}_{t_{k-1\vee 0}}\right)h + \sigma\left(t_k, \tilde{X}^{i,N,M}_{t_k}, \tilde{\mu}^{i,N,M}_{t_k}\right)\Delta W^i_{t_k},$$

where $\hat{X}_0^{i,N,M} = X_0^i$ and $\vee$ denotes the maximum. The scheme's continuous time version is

$$\hat{X}_t^{i,N,M} = X_0^i + \int_0^t b\left((\kappa(s) - h) \vee 0, \tilde{X}_{\kappa(s)}^{i,N,M}, \tilde{\mu}_{(\kappa(s)-h)\vee 0}^{X,N,M}\right) \mathrm{d}s \\ + \int_0^t \sigma\left(\kappa(s), \tilde{X}_{\kappa(s)}^{i,N,M}, \tilde{\mu}_{\kappa(s)}^{i,N,M}\right) \mathrm{d}W_s^i. \tag{9.3.11}$$

The first result we wish to present is that the discrete and continuous versions stay close to one another, up to the stopping time (9.3.10).

Lemma 9.3.9. *Let Assumption 8.2.1, 9.1.1 and H1 (in Assumption 9.1.6) hold and fix $h^* < 1/\max(L_b, 2\beta)$. Further assume $X_0 \in L^{4(q+1)}(\mathbb{R}^d)$. Then for $1 \le p \le 4$ the following holds for $0 < h \le h^*$,*

$$\sup_{1\le i\le N} \sup_{0\le k\le M} \mathbb{E}\big[|\hat{X}_{t_k}^{i,N,M} - \tilde{X}_{t_k}^{i,N,M}|^p \mathbb{1}_{\{k\le\lambda_m^i\}}\big] \le C(m,p)h^p\,.$$

Moreover, we also have the following relation between $\hat{X}$ and F for all $1 \le k \le M$,

$$|\hat{X}_{t_k}^{i,N,M}|^2 \ge \frac{1}{2}|F(t_{k-1}, \tilde{X}_{t_k}^{i,N,M}, \tilde{\mu}_{t_{k-1}}^{X,N,M})|^2 - |b(t_0, \tilde{X}_{t_0}^{i,N,M}, \tilde{\mu}_{t_0}^{X,N,M})h|^2\,. \tag{9.3.12}$$

Proof. To show the first part we start by noting the following useful relation between (9.0.2) and (9.3.11), namely for $1 \le k \le M$,

$$\hat{X}_{t_k}^{i,N,M} - \tilde{X}_{t_k}^{i,N,M} = \big(b(t_0, \tilde{X}_{t_0}^{i,N,M}, \tilde{\mu}_{t_0}^{X,N,M}) - b(t_{k-1}, \tilde{X}_{t_k}^{i,N,M}, \tilde{\mu}_{t_{k-1}}^{X,N,M})\big)h\,.$$

Noting that one can bound,

$$|b(t_0, \tilde{X}_{t_0}^{i,N,M}, \tilde{\mu}_{t_0}^{X,N,M}) - b(t_{k-1}, \tilde{X}_{t_k}^{i,N,M}, \tilde{\mu}_{t_{k-1}}^{X,N,M})| \\ \le C\big(1 + |t_k|^{1/2} + |\tilde{X}_{t_0}^{i,N,M}|^{q+1} + |\tilde{X}_{t_k}^{i,N,M}|^{q+1}\big)\,,$$

where we have used the growth bounds on the coefficient b, in particular Assumption H1. Hence,

$$\mathbb{E}\big[|\hat{X}_{t_k}^{i,N,M} - \tilde{X}_{t_k}^{i,N,M}|^p \mathbb{1}_{\{k\le\lambda_m^i\}}\big] \\ \le C(p)h^p\big(1 + |t_k|^{p/2} + \mathbb{E}\big[|\tilde{X}_{t_0}^{i,N,M}|^{p(q+1)}\mathbb{1}_{\{k\le\lambda_m^i\}}\big] + \mathbb{E}\big[|\tilde{X}_{t_k}^{i,N,M}|^{p(q+1)}\mathbb{1}_{\{k\le\lambda_m^i\}}\big]\big)\,.$$

One observes that the terms on the RHS are bounded by $C(p,m)$ for $p \le 4$ since $X_0 \in L^{4(q+1)}(\mathbb{R}^d)$ and Lemma 9.3.7. This completes the first part of the proof.

For the second part, recall from the relation between (9.0.2) and (9.3.11), one has,

$$\hat{X}_{t_k}^{i,N,M} = b(t_0, \tilde{X}_{t_0}^{i,N,M}, \tilde{\mu}_{t_0}^{X,N,M})h + \tilde{X}_{t_k}^{i,N,M} - b(t_{k-1}, \tilde{X}_{t_k}^{i,N,M}, \tilde{\mu}_{t_{k-1}}^{X,N,M})h \\ = b(t_0, \tilde{X}_{t_0}^{i,N,M}, \tilde{\mu}_{t_0}^{X,N,M})h + F(t_{k-1}, \tilde{X}_{t_k}^{i,N,M}, \tilde{\mu}_{t_{k-1}}^{X,N,M})\,.$$

Using the reverse triangle inequality we obtain,

$$|\hat{X}_{t_k}^{i,N,M}|^2 \ge -|b(t_0, \tilde{X}_{t_0}^{i,N,M}, \tilde{\mu}_{t_0}^{X,N,M})h| + |F(t_{k-1}, \tilde{X}_{t_k}^{i,N,M}, \tilde{\mu}_{t_{k-1}}^{X,N,M})|\,.$$

The result follows from squaring both sides and applying the generalisation of Young's inequality, namely,

$$|b(t_0, \tilde{X}_{t_0}^{i,N,M}, \tilde{\mu}_{t_0}^{X,N,M})h||F(t_{k-1}, \tilde{X}_{t_k}^{i,N,M}, \tilde{\mu}_{t_{k-1}}^{X,N,M})| \\ \le |b(t_0, \tilde{X}_{t_0}^{i,N,M}, \tilde{\mu}_{t_0}^{X,N,M})h|^2 + \frac{1}{4}|F(t_{k-1}, \tilde{X}_{t_k}^{i,N,M}, \tilde{\mu}_{t_{k-1}}^{X,N,M})|^2\,.$$

□

The next result we wish to present is that both schemes do not blow up in finite time, for this we define a new stopping time,

$$\eta^i_m := \inf\big\{t \geq 0 : |\hat{X}^{i,N,M}_t| \geq m\,, \text{ or } |\tilde{X}^{i,N,M}_{\kappa(t)}| > m\big\}\,.$$

Lemma 9.3.10. *Let Assumption 8.2.1, 9.1.1 and H1 (in Assumption 9.1.6) hold, fix $h^* < 1/\max(L_b, 2\beta)$ and assume $X_0 \in L^{4(q+1)}(\mathbb{R}^d)$. Then, for any $\epsilon > 0$, there exists a m^* such that, for any $m \geq m^*$ we can find a $h^*_0(m)$ (note the dependence on m) so that,*

$$\sup_{1\leq i\leq N} \mathbb{P}(\eta^i_m < T) \leq \epsilon\,, \text{ for any } 0 < h \leq h^*_0(m).$$

Proof. Note due to the initial condition being random we must be careful with how we set m, we shall come back to this later. Let us start by applying Itô to the stopped version of (9.3.11),

$$\begin{aligned}|\hat{X}^{i,N,M}_{T\wedge\eta^i_m}|^2 = |X^i_0|^2 + \int_0^{T\wedge\eta^i_m} 2\langle \hat{X}^{i,N,M}_t, b\big((\kappa(s)-h)\vee 0, \tilde{X}^{i,N,M}_{\kappa(s)}, \tilde{\mu}^{X,N,M}_{(\kappa(s)-h)\vee 0}\big)\rangle \\ + \sum_{a=1}^{l} |\sigma_a(\kappa(s), \tilde{X}^{i,N,M}_{\kappa(s)}, \tilde{\mu}^{i,N,M}_{\kappa(s)})|^2 \mathrm{d}s + \int_0^{T\wedge\eta^i_m} \langle \hat{X}^{i,N,M}_t, \sigma(\kappa(s), \tilde{X}^{i,N,M}_{\kappa(s)}, \tilde{\mu}^{i,N,M}_{\kappa(s)})\mathrm{d}W^i_s\rangle.\end{aligned}$$

We now look to bound the various integrands, firstly one can observe

$$\begin{aligned}&\langle \hat{X}^{i,N,M}_t, b\big((\kappa(s)-h)\vee 0, \tilde{X}^{i,N,M}_{\kappa(s)}, \tilde{\mu}^{X,N,M}_{(\kappa(s)-h)\vee 0}\big)\rangle + \sum_{a=1}^{l} |\sigma_a\big(\kappa(s), \tilde{X}^{i,N,M}_{\kappa(s)}, \tilde{\mu}^{i,N,M}_{\kappa(s)}\big)|^2 \\ &= \langle \hat{X}^{i,N,M}_t - \tilde{X}^{i,N,M}_{\kappa(s)}, b\big((\kappa(s)-h)\vee 0, \tilde{X}^{i,N,M}_{\kappa(s)}, \tilde{\mu}^{X,N,M}_{(\kappa(s)-h)\vee 0}\big)\rangle \\ &\quad + \langle \tilde{X}^{i,N,M}_{\kappa(s)}, b\big((\kappa(s)-h)\vee 0, \tilde{X}^{i,N,M}_{\kappa(s)}, \tilde{\mu}^{X,N,M}_{(\kappa(s)-h)\vee 0}\big)\rangle + \sum_{a=1}^{l} |\sigma_a\big(\kappa(s), \tilde{X}^{i,N,M}_{\kappa(s)}, \tilde{\mu}^{i,N,M}_{\kappa(s)}\big)|^2 \\ &\leq C|\hat{X}^{i,N,M}_t - \tilde{X}^{i,N,M}_{\kappa(s)}|(1 + |\tilde{X}^{i,N,M}_{\kappa(s)}|^{q+1}) + \alpha + \beta|\tilde{X}^{i,N,M}_{\kappa(s)}|^2\,,\end{aligned}$$

where we used Cauchy-Schwarz, polynomial growth bound and monotone growth to obtain the final inequality.

Taking expectations and noting that due to the stopping time the stochastic integral is a martingale i.e. has second moment, we obtain,

$$\begin{aligned}&\mathbb{E}[|\hat{X}^{i,N,M}_{T\wedge\eta^i_m}|^2] \\ &\leq \mathbb{E}[|X^i_0|^2] + \mathbb{E}\Big[\int_0^{T\wedge\eta^i_m} C|\hat{X}^{i,N,M}_s - \tilde{X}^{i,N,M}_{\kappa(s)}|(1 + |\tilde{X}^{i,N,M}_{\kappa(s)}|^{q+1}) + \alpha + \beta|\tilde{X}^{i,N,M}_{\kappa(s)}|^2 \mathrm{d}s\Big].\end{aligned}$$

To proceed we note the following, $|\tilde{X}^{i,N,M}_{\kappa(s)}|^2 \leq 2(|\tilde{X}^{i,N,M}_{\kappa(s)} - \hat{X}^{i,N,M}_s|^2 + |\hat{X}^{i,N,M}_s|^2)$ and also

$$\int_0^{T\wedge\eta^i_m} |\hat{X}^{i,N,M}_s - \tilde{X}^{i,N,M}_{\kappa(s)}|^2 \mathrm{d}s \leq C(m) \int_0^{T\wedge\eta^i_m} |\hat{X}^{i,N,M}_s - \tilde{X}^{i,N,M}_{\kappa(s)}| \mathrm{d}s\,,$$

where we used the fact that the stopping time ensures $\tilde{X}$ and $\hat{X}$ are $\leq m$ for $s < \eta^i_m$ and $s = \eta^i_m$ has measure zero. The same reasoning also implies,

$$\int_0^{T\wedge\eta^i_m} C|\hat{X}^{i,N,M}_s - \tilde{X}^{i,N,M}_{\kappa(s)}|(1 + |\tilde{X}^{i,N,M}_{\kappa(s)}|^{q+1})\mathrm{d}s \leq C(m) \int_0^{T\wedge\eta^i_m} |\hat{X}^{i,N,M}_s - \tilde{X}^{i,N,M}_{\kappa(s)}| \mathrm{d}s.$$

Hence the following result holds,

$$\mathbb{E}[|\hat{X}^{i,N,M}_{T\wedge\eta^i_m}|^2] \leq \mathbb{E}[|X^i_0|^2] + C\mathbb{E}\Big[\int_0^{T\wedge\eta^i_m} C(m)|\hat{X}^{i,N,M}_s - \tilde{X}^{i,N,M}_{\kappa(s)}| + 1 + \beta|\hat{X}^{i,N,M}_s|^2 \mathrm{d}s\Big]\,.$$

The next step is of course to take the expectation inside the integral, let us start by noting the difference term can be bounded as,

$$
\begin{aligned}
&\mathbb{E}\Big[\int_0^{T\wedge\eta_m^i}|\hat{X}_s^{i,N,M}-\tilde{X}_{\kappa(s)}^{i,N,M}|\mathrm{d}s\Big]\\
&\qquad\leq \mathbb{E}\Big[\int_0^{T\wedge\eta_m^i}|\hat{X}_s^{i,N,M}-\hat{X}_{\kappa(s)}^{i,N,M}|\mathrm{d}s+\int_0^{T\wedge\eta_m^i}|\hat{X}_{\kappa(s)}^{i,N,M}-\tilde{X}_{\kappa(s)}^{i,N,M}|\mathrm{d}s\Big]\\
&\qquad\leq \mathbb{E}\Big[h\int_0^{T\wedge\eta_m^i}|b\left((\kappa(s)-h)\vee 0,\tilde{X}_{\kappa(s)}^{i,N,M},\tilde{\mu}_{(\kappa(s)-h)\vee 0}^{X,N,M}\right)|\mathrm{d}s\Big]\\
&\qquad\quad+\mathbb{E}\Big[\int_0^{T\wedge\eta_m^i}|\sigma\left(\kappa(s),\tilde{X}_{\kappa(s)}^{i,N,M},\tilde{\mu}_{\kappa(s)}^{i,N,M}\right)(W_s^i-W_{\kappa(s)}^i)|\mathrm{d}s\Big]+C(m)h^{1/2},
\end{aligned}
$$

where we have used Lemma 9.3.9 for the final inequality. For the other terms, one can note due to the growth assumptions on b, that,

$$
\mathbb{E}\Big[h\int_0^{T\wedge\eta_m^i}|b\left((\kappa(s)-h)\vee 0,\tilde{X}_{\kappa(s)}^{i,N,M},\tilde{\mu}_{(\kappa(s)-h)\vee 0}^{X,N,M}\right)|\mathrm{d}s\Big]\leq C(m)\,h.
$$

The term involving σ is more complex, however, we can bound as follows,

$$
\begin{aligned}
&\mathbb{E}\Big[\int_0^{T\wedge\eta_m^i}|\sigma\left(\kappa(s),\tilde{X}_{\kappa(s)}^{i,N,M},\tilde{\mu}_{\kappa(s)}^{i,N,M}\right)(W_s^i-W_{\kappa(s)}^i)|\mathrm{d}s\Big]\\
&\leq C\int_0^T\sum_{a=1}^{l}\mathbb{E}\Big[|\sigma_a\left(\kappa(s),\tilde{X}_{\kappa(s)}^{i,N,M},\tilde{\mu}_{\kappa(s)}^{i,N,M}\right)||(W_s^i-W_{\kappa(s)}^i)_a|\mathbb{1}_{\{\kappa(s)\leq t_{\lambda_m^i}\}}\Big]\mathrm{d}s\\
&\leq C\int_0^T\sum_{a=1}^{l}h^{1/2}(1+\mathbb{E}[|\tilde{X}_{\kappa(s)\wedge t_{\lambda_m^i}}^{i,N,M}|^2])\mathrm{d}s\leq C(m)h^{1/2},
\end{aligned}
$$

where the bound follows from Lemma 9.3.7. Further, since $|\hat{X}_s^{i,N,M}|\geq 0$, we obtain,

$$
\mathbb{E}\Big[\int_0^{T\wedge\eta_m^i}|\hat{X}_s^{i,N,M}|^2\mathrm{d}s\Big]\leq\int_0^T\mathbb{E}\big[|\hat{X}_{s\wedge\eta_m^i}^{i,N,M}|^2\big]\mathrm{d}s.
$$

Hence,

$$
\begin{aligned}
\mathbb{E}[|\hat{X}_{T\wedge\eta_m^i}^{i,N,M}|^2]\leq&\mathbb{E}[|X_0^i|^2]+C(m)h^{1/2}+C\int_0^T 1+\beta\mathbb{E}\big[|\hat{X}_{s\wedge\eta_m^i}^{i,N,M}|^2\big]\mathrm{d}s\\
\leq&\big(\mathbb{E}[|X_0^i|^2]+C+C(m)h^{1/2}\big)\exp(C\beta T)\,,
\end{aligned}
\tag{9.3.13}
$$

where the final inequality follows from Grönwall.

In order to obtain an upper bound on the probability of the stopping time occurring we look to obtain a lower bound for (9.3.11) at the stopping time. For the moment let us take $X_0^i<m$, hence $\eta_m^i>0$, there are now two possible ways the stopping time can be reached, if $\hat{X}$ hits the boundary first then we have $|\hat{X}_{\eta_m^i}^{i,N,M}|=m$ and if $\tilde{X}$ hits the boundary first we have $|\tilde{X}_{\eta_m^i}^{i,N,M}|>m$.

In the case that $\hat{X}$ hits the boundary first, the lower bound is obvious, namely $|\hat{X}_{\eta_m^i}^{i,N,M}|=m$. For the second case it is less obvious. Recalling (9.3.12) and Lemma 9.3.6 we obtain the following lower bound for,

$$
|\hat{X}_{t_k}^{i,N,M}|^2\geq\frac{1}{2}\big((1-2h\beta)|\tilde{X}_{t_k}^{i,N,M}|^2-2h\alpha\big)-|b(t_0,\tilde{X}_{t_0}^{i,N,M},\tilde{\mu}_{t_0}^{X,N,M})h|^2\,,
$$

where again we are taking $k\geq 1$ here, but this is not a problem since we are assuming for the moment $X_0^i<m$. Observing that this lower bound holds independent of which process triggers

the stopping condition we can say w.l.o.g. that,

$$
\begin{aligned}
m^2 &\geq |\hat{X}^{i,N,M}_{\eta^i_m}|^2 \mathbb{1}_{\{|X^i_0|<m\}} \\
&\geq \frac{1}{2}\big((1-2h\beta)m^2 - 2h\alpha\big)\mathbb{1}_{\{|X^i_0|<m\}} - |b(t_0, \tilde{X}^{i,N,M}_{t_0}, \tilde{\mu}^{X,N,M}_{t_0})h|^2 \mathbb{1}_{\{|X^i_0|<m\}} \,.
\end{aligned}
$$

Therefore,

$$
|\hat{X}^{i,N,M}_{\eta^i_m}|^2 \mathbb{1}_{\{|X^i_0|<m\}} \geq (C_1 m^2 - C_2 h)\mathbb{1}_{\{|X^i_0|<m\}} - C(m)h^2 \mathbb{1}_{\{|X^i_0|<m\}} \,,
$$

where $|b(t_0, \tilde{X}^{i,N,M}_{t_0}, \tilde{\mu}^{X,N,M}_{t_0})|\mathbb{1}_{\{|X^i_0|<m\}} \leq C(m)\mathbb{1}_{\{|X^i_0|<m\}}$ via the growth condition on b. Let us now combine these results to obtain an upper bound for the probability of the stopping time, notice that,

$$
\begin{aligned}
\mathbb{E}[|\hat{X}^{i,N,M}_{T\wedge\eta^i_m}|^2] &\geq \mathbb{E}[|X^i_0|^2 \mathbb{1}_{\{|X^i_0|\geq m\}}] + \mathbb{E}[|\hat{X}^{i,N,M}_{\eta^i_m}|^2 \mathbb{1}_{\{0<\eta^i_m<T\}}] \\
&\geq \mathbb{P}(\eta^i_m = 0) + \big((C_1 m^2 - C_2 h) - C(m)h^2\big)\, \mathbb{P}(\{|X^i_0| < m\} \cap \{0 < \eta^i_m < T\}) \,.
\end{aligned}
$$

Leaving the second term for the moment, and noting that X^i_0 is uniformly integrable, then for any $\epsilon > 0$ one obtains,

$$
\mathbb{P}(\eta^i_m = 0) \leq m\mathbb{P}(|X^i_0| \geq m) \leq \mathbb{E}[|X^i_0|\mathbb{1}_{\{|X^i_0|\geq m\}}] \leq \frac{\epsilon}{3} \,,
$$

for m sufficiently large, call this point m^*. It is also useful to note that $\mathbb{P}(\{|X^i_0| < m\} \cap \{0 < \eta^i_m < T\}) = \mathbb{P}(\{0 < \eta^i_m < T\})$. It is clear from our previous analysis that for m large enough and (9.3.13) the probability can be bounded by,

$$
\mathbb{P}(0 < \eta^i_m < T) \leq \frac{\mathbb{E}[|\hat{X}^{i,N,M}_{T\wedge\eta^i_m}|^2]}{(C_1 m^2 - C_2 h - C(m)h^2)} \leq \frac{\big(\mathbb{E}[|X^i_0|^2] + C + C(m)h^{1/2}\big)\exp(C\beta T)}{C_1 m^2 - C_2 h - C(m)h^2} \,.
$$

Now the goal is to bound this by $2\epsilon/3$, we already have taken m sufficiently large to obtain the last inequality, now consider for any given m, $h^*_{01}(m) :\ C_2 h^*_{01}(m) + C(m)h^*_{01}(m)^2 \leq 1$. It is clear for $0 < h < h^*_{01}(m)$ the same bound holds. Then for the same ϵ as before choose m large enough such that,

$$
\frac{\big(\mathbb{E}[|X^i_0|^2] + C\big)\exp(C\beta T)}{C_1 m^2 - 1} \leq \frac{\epsilon}{3} \,.
$$

Redefine m^* as the corresponding maximum of this m and m^*. Now for any $m \geq m^*$, define $h^*_{02}(m)$ such that,

$$
\frac{C(m)(h^*_{02})^{1/2}\exp(C\beta T)}{C_1 m^2 - 1} \leq \frac{\epsilon}{3} \,.
$$

Again for $0 < h < h^*_{02}(m)$ the above inequality holds. Hence for any $m \geq m^*$ and any $0 < h < \min(h^*_{01}(m), h^*_{02}(m))$, we have, $\mathbb{P}(\eta^i_m < T) \leq \mathbb{P}(\eta^i_m = 0) + \mathbb{P}(0 < \eta^i_m < T) \leq \epsilon$. □

We now look towards showing our strong convergence result, firstly by showing convergence between (9.3.11) and (8.2.2) and then (9.0.2) and (8.2.2). From this point onwards we require H2 (in Assumption 9.1.6).

Remark 9.3.11 (On the diffusion coefficient σ being independent of the measure)**.** *The reason we cannot allow σ to have measure dependence is because our stopping time arguments do not work. Namely in order for two diffusion coefficients to be similar we require all N particles to be close to one another, not just the ith particle. As it turns out though, this is not a problem for the drift term, so we make no change to the measure dependence there.*

Recalling the stopping time in Proposition 9.3.5, we now define $\theta^i_m := \tau^i_m \wedge \eta^i_m$ and have the following convergence result.

Lemma 9.3.12. *Let Assumption 8.2.1, 9.1.1, the full Assumption 9.1.6 hold, fix $h^* < 1/\max(L_b, 2\beta)$ and assume $X_0 \in L^{4(q+1)}(\mathbb{R}^d)$. Then*

$$\sup_{1\leq i\leq N} \mathbb{E}[\sup_{0\leq t\leq T} |\hat{X}^{i,N,M}_{t\wedge\theta^i_m} - X^{i,N}_{t\wedge\theta^i_m}|^2] \leq C(m)h\,.$$

Proof. For ease of presentation we denote by $\overline{\kappa}(s) := (\kappa(s)-h)\vee 0$. As is standard we start by applying Itô to the difference to obtain,

$$\begin{aligned}
&|X^{i,N}_{t\wedge\theta^i_m} - \hat{X}^{i,N,M}_{t\wedge\theta^i_m}|^2 \\
&= \int_0^{t\wedge\theta^i_m} 2\langle X^{i,N}_s - \hat{X}^{i,N,M}_s, b(s, X^{i,N}_s, \mu^{X,N}_s) - b(\overline{\kappa}(s), \tilde{X}^{i,N,M}_{\kappa(s)}, \mu^{X,N,M}_{(\kappa(s)-h)})\rangle \\
&\qquad\qquad + \sum_{a=1}^{l} |\sigma(s, X^{i,N}_s) - \sigma(\kappa(s), \tilde{X}^{i,N,M}_{\kappa(s)})|^2 \mathrm{d}s \\
&+ \int_0^{t\wedge\theta^i_m} 2\langle X^{i,N}_s - \hat{X}^{i,N,M}_s, \big(\sigma(s, X^{i,N}_s) - \sigma(\kappa(s), \tilde{X}^{i,N,M}_{\kappa(s)})\big)\mathrm{d}W^i_s\rangle\,.
\end{aligned}$$

By writing out the drift term we have that,

$$\begin{aligned}
&\langle X^{i,N}_s - \hat{X}^{i,N,M}_s, b(s, X^{i,N}_s, \mu^{X,N}_s) - b(\overline{\kappa}(s), \tilde{X}^{i,N,M}_{\kappa(s)}, \mu^{X,N,M}_{(\kappa(s)-h)})\rangle \\
&= \langle X^{i,N}_s - \hat{X}^{i,N,M}_s, b(s, X^{i,N}_s, \mu^{X,N}_s) - b(s, \hat{X}^{i,N,M}_s, \mu^{X,N}_s)\rangle \\
&\ + \langle X^{i,N}_s - \hat{X}^{i,N,M}_s, b(s, \hat{X}^{i,N,M}_s, \mu^{X,N}_s) - b(\overline{\kappa}(s), \hat{X}^{i,N,M}_s, \mu^{X,N}_s)\rangle \\
&\ + \langle X^{i,N}_s - \hat{X}^{i,N,M}_s, b(\overline{\kappa}(s), \hat{X}^{i,N,M}_s, \mu^{X,N}_s) - b(\overline{\kappa}(s), \hat{X}^{i,N,M}_{\kappa(s)}, \mu^{X,N}_s)\rangle \\
&\ + \langle X^{i,N}_s - \hat{X}^{i,N,M}_s, b(\overline{\kappa}(s), \hat{X}^{i,N,M}_{\kappa(s)}, \mu^{X,N}_s) - b(\overline{\kappa}(s), \tilde{X}^{i,N,M}_{\kappa(s)}, \mu^{X,N}_s)\rangle \\
&\ + \langle X^{i,N}_s - \hat{X}^{i,N,M}_s, b(\overline{\kappa}(s), \tilde{X}^{i,N,M}_{\kappa(s)}, \mu^{X,N}_s) - b(\overline{\kappa}(s), \tilde{X}^{i,N,M}_{\kappa(s)}, \mu^{X,N,M}_{(\kappa(s)-h)})\rangle \\
&\leq C\Big(|X^{i,N}_s - \hat{X}^{i,N,M}_s|^2 + |X^{i,N}_s - \hat{X}^{i,N,M}_s| + h \\
&\qquad + (1 + |\hat{X}^{i,N,M}_s|^{2q} + |\hat{X}^{i,N,M}_{\kappa(s)}|^{2q})|\hat{X}^{i,N,M}_s - \hat{X}^{i,N,M}_{\kappa(s)}|^2 \\
&\qquad\qquad + (1 + |\hat{X}^{i,N,M}_{\kappa(s)}|^{2q} + |\tilde{X}^{i,N,M}_{\kappa(s)}|^{2q})|\hat{X}^{i,N,M}_{\kappa(s)} - \tilde{X}^{i,N,M}_{\kappa(s)}|^2\Big)\,,
\end{aligned}$$

where we have used the growth bounds on b (in particular bounded in measure) along with several applications of Cauchy-Schwarz and Young's inequality. Similar arguments yield the following bound for the diffusion,

$$\begin{aligned}
&|\sigma(s, X^{i,N}_s) - \sigma(\kappa(s), \tilde{X}^{i,N,M}_{\kappa(s)})| \\
&\qquad \leq C(h^{1/2} + |X^{i,N}_s - \hat{X}^{i,N,M}_s| + |\hat{X}^{i,N,M}_s - \hat{X}^{i,N,M}_{\kappa(s)}| + |\hat{X}^{i,N,M}_{\kappa(s)} - \tilde{X}^{i,N,M}_{\kappa(s)}|).
\end{aligned}$$

Ultimately we need to take supremum and expected values, hence we wish to bound

$$\mathbb{E}\Big[\sup_{0\leq r\leq t\wedge\theta^i_m} \int_0^r 2\langle X^{i,N}_s - \hat{X}^{i,N,M}_s, \big(\sigma(s, X^{i,N}_s) - \sigma(\kappa(s), \tilde{X}^{i,N,M}_{\kappa(s)})\big)\mathrm{d}W^i_s\rangle\Big]\,.$$

We use Burkholder Davis Gundy inequality, however care is needed since the terminal time is a stopping time. It turns out the usual upper bound still holds (see for example [Pro05, pg. 226]), hence we obtain by using Young's inequality,

$$\begin{aligned}
&\mathbb{E}\Big[\sup_{0\leq r\leq t\wedge\theta^i_m}\int_0^r 2\langle X^{i,N}_s-\hat{X}^{i,N,M}_s,\big(\sigma(s,X^{i,N}_s)-\sigma(\kappa(s),\tilde{X}^{i,N,M}_{\kappa(s)})\big)\mathrm{d}W^i_s\rangle\Big]\\
&\leq C\mathbb{E}\Big[\Big(\int_0^{t\wedge\theta^i_m}|X^{i,N}_s-\hat{X}^{i,N,M}_s|^2\sum_{a=1}^{l}|\sigma_a(s,X^{i,N}_s)-\sigma_a(\kappa(s),\tilde{X}^{i,N,M}_{\kappa(s)})|^2\mathrm{d}s\Big)^{1/2}\Big]\\
&\leq \frac{1}{2}\mathbb{E}\Big[\sup_{0\leq s\leq t\wedge\theta^i_m}|X^{i,N}_s-\hat{X}^{i,N,M}_s|^2\Big]+C\mathbb{E}\Big[\int_0^{t\wedge\theta^i_m}\sum_{a=1}^{l}|\sigma_a(s,X^{i,N}_s)-\sigma_a(\kappa(s),\tilde{X}^{i,N,M}_{\kappa(s)})|^2\mathrm{d}s\Big]\,.
\end{aligned}$$

Taking supremum and expectations of our original difference and using these bounds we obtain the following inequality

$$\begin{aligned}
&\frac{1}{2}\mathbb{E}\left[\sup_{0\leq t\leq T\wedge\theta^i_m}|X^{i,N}_{t\wedge\theta^i_m}-\hat{X}^{i,N,M}_{t\wedge\theta^i_m}|^2\right]\\
&\leq \mathbb{E}\Big[\int_0^{T\wedge\theta^i_m}C\big(|X^{i,N}_s-\hat{X}^{i,N,M}_s|^2+|X^{i,N}_s-\hat{X}^{i,N,M}_s|\\
&+(1+|\hat{X}^{i,N,M}_s|^{2q}+|\hat{X}^{i,N,M}_{\kappa(s)}|^{2q})|\hat{X}^{i,N,M}_s-\hat{X}^{i,N,M}_{\kappa(s)}|^2\\
&+h+(1+|\hat{X}^{i,N,M}_{\kappa(s)}|^{2q}+|\tilde{X}^{i,N,M}_{\kappa(s)}|^{2q})|\hat{X}^{i,N,M}_{\kappa(s)}-\tilde{X}^{i,N,M}_{\kappa(s)}|^2\big)\\
&+C\sum_{a=1}^{l}\Big(h+|X^{i,N}_s-\hat{X}^{i,N,M}_s|^2+|\hat{X}^{i,N,M}_s-\hat{X}^{i,N,M}_{\kappa(s)}|^2+|\hat{X}^{i,N,M}_{\kappa(s)}-\tilde{X}^{i,N,M}_{\kappa(s)}|^2\Big)\,\mathrm{d}s\Big]\,.
\end{aligned}$$

The goal is to use a Grönwall type inequality, hence we want to bring the expectation inside the integral, collecting common terms and arguing as previous we obtain,

$$\begin{aligned}
&\mathbb{E}\left[\sup_{0\leq t\leq T\wedge\theta^i_m}|X^{i,N}_{t\wedge\theta^i_m}-\hat{X}^{i,N,M}_{t\wedge\theta^i_m}|^2\right]\\
&\leq C\big(hT+\int_0^T\mathbb{E}\Big[\sup_{0\leq r\leq s}|X^{i,N}_{r\wedge\theta^i_m}-\hat{X}^{i,N,M}_{r\wedge\theta^i_m}|^2\Big]+\mathbb{E}\Big[\sup_{0\leq r\leq s}|X^{i,N}_{r\wedge\theta^i_m}-\hat{X}^{i,N,M}_{r\wedge\theta^i_m}|\Big]\\
&\quad+\mathbb{E}\Big[(1+|\hat{X}^{i,N,M}_s|^{2q}+|\hat{X}^{i,N,M}_{\kappa(s)}|^{2q})|\hat{X}^{i,N,M}_s-\hat{X}^{i,N,M}_{\kappa(s)}|^2\mathbb{1}_{\{s\leq\theta^i_m\}}\Big]\\
&\quad+\mathbb{E}\Big[(1+|\hat{X}^{i,N,M}_{\kappa(s)}|^{2q}+|\tilde{X}^{i,N,M}_{\kappa(s)}|^{2q})|\hat{X}^{i,N,M}_{\kappa(s)}-\tilde{X}^{i,N,M}_{\kappa(s)}|^2\mathbb{1}_{\{s\leq\theta^i_m\}}\Big]\mathrm{d}s\big)\,.
\end{aligned}$$

Noting $\mathbb{1}_{\{\cdot\}}=\mathbb{1}^2_{\{\cdot\}}$, we obtain via Cauchy-Schwarz inequality,

$$\begin{aligned}
&\mathbb{E}\Big[(1+|\hat{X}^{i,N,M}_s|^{2q}+|\hat{X}^{i,N,M}_{\kappa(s)}|^{2q})|\hat{X}^{i,N,M}_s-\hat{X}^{i,N,M}_{\kappa(s)}|^2\mathbb{1}_{\{s\leq\theta^i_m\}}\Big]\\
&\qquad\qquad\leq C(m)\mathbb{E}\Big[|\hat{X}^{i,N,M}_s-\hat{X}^{i,N,M}_{\kappa(s)}|^4\mathbb{1}_{\{s\leq\theta^i_m\}}\Big].
\end{aligned}$$

Noting that

$$|\hat{X}^{i,N,M}_s-\hat{X}^{i,N,M}_{\kappa(s)}|\leq|b\left(\overline{\kappa}(s),\tilde{X}^{i,N,M}_{\kappa(s)},\tilde{\mu}^{X,N,M}_{(\kappa(s)-h)\vee 0}\right)|h+|\sigma\left(\kappa(s),\tilde{X}^{i,N,M}_{\kappa(s)}\right)(W^i_s-W^i_{\kappa(s)})|\,,$$

which implies,

$$\begin{aligned}
&\mathbb{E}\Big[|\hat{X}^{i,N,M}_s-\hat{X}^{i,N,M}_{\kappa(s)}|^4\mathbb{1}_{\{s\leq\theta^i_m\}}\Big]\\
&\qquad\leq Ch^4\mathbb{E}\Big[(1+|\tilde{X}^{i,N,M}_{\kappa(s)}|^{4(q+1)})\mathbb{1}_{\{s\leq\theta^i_m\}}\Big]\\
&\qquad\quad+C\mathbb{E}\Big[(1+|\tilde{X}^{i,N,M}_{\kappa(s)}|^4)\mathbb{1}_{\{s\leq\theta^i_m\}}\Big]\mathbb{E}\Big[(W^i_s-W^i_{\kappa(s)})^4\Big]\leq C(m)h^2\,,
\end{aligned}$$

where we used Lemma 9.3.7 to obtain the final inequality (note by assumption $X_0 \in L^{4(q+1)}(\mathbb{R}^d)$). Arguing in the exact same fashion also yields,

$$\mathbb{E}\Big[(1+|\hat{X}^{i,N,M}_{\kappa(s)}|^{2q}+|\tilde{X}^{i,N,M}_{\kappa(s)}|^{2q})|\hat{X}^{i,N,M}_{\kappa(s)}-\tilde{X}^{i,N,M}_{\kappa(s)}|^2\mathbb{1}_{\{s\leq\theta^i_m\}}\Big]\leq C(m)h^2\,.$$

Substituting these bounds then implies,

$$\begin{aligned}&\mathbb{E}\left[\sup_{0\leq t\leq T\wedge\theta^i_m}|X^{i,N}_{t\wedge\theta^i_m}-\hat{X}^{i,N,M}_{t\wedge\theta^i_m}|^2\right]\\&\qquad\leq C(m)h+C\int_0^T\mathbb{E}\Big[\sup_{0\leq r\leq s}|X^{i,N}_{r\wedge\theta^i_m}-\hat{X}^{i,N,M}_{r\wedge\theta^i_m}|^2\Big]+\mathbb{E}\Big[\sup_{0\leq r\leq s}|X^{i,N}_{r\wedge\theta^i_m}-\hat{X}^{i,N,M}_{r\wedge\theta^i_m}|\Big]\mathrm{d}s\,.\end{aligned}$$

Observing for a random variable Y, $\mathbb{E}[|Y|]\leq\mathbb{E}[|Y|^2]^{1/2}$, we observe that the above inequality can be tackled using Perov's inequality (a nonlinear version of Grönwall, see [MPF12, pg. 360]), hence,

$$\mathbb{E}\Big[\sup_{0\leq t\leq T\wedge\theta^i_m}|X^{i,N}_{t\wedge\theta^i_m}-\hat{X}^{i,N,M}_{t\wedge\theta^i_m}|^2\Big]\leq\Big(C(m)^{1/2}h^{1/2}\exp\Big(CT+C\int_0^T\exp\big(Ct\big)\mathrm{d}t\Big)\Big)^2\leq C(m)h\,,$$

which gives the result we set out to show. □

We now can prove our main implicit scheme result.

Proposition 9.1.8. Let us define the error term as $E_r(T)^i = X^{i,N}_T-\tilde{X}^{i,N,M}_T$ and also let us note a more general version of Young's inequality,

$$x^s y\leq\frac{\delta s}{2}x^2+\frac{2-s}{2\delta^{s/(2-s)}}y^{2/(2-s)}\,,\quad\forall\ x,\ y,\ \delta>0\,.$$

Hence,

$$\begin{aligned}\mathbb{E}[|X^{i,N}_T-\tilde{X}^{i,N,M}_T|^s]\leq 2^{s-1}\big(&\mathbb{E}[|X^{i,N}_T-\hat{X}^{i,N,M}_T|^s\mathbb{1}_{\{\tau^i_m>T,\ \eta^i_m>T\}}]\\&+\mathbb{E}[|\hat{X}^{i,N,M}_T-\tilde{X}^{i,N,M}_T|^s\mathbb{1}_{\{\tau^i_m>T,\ \eta^i_m>T\}}]\big)\\&+\frac{\delta s}{2}\mathbb{E}[|E_r(T)^i|^2]+\frac{2-s}{2\delta^{s/(2-s)}}\mathbb{E}[\mathbb{1}_{\{\tau^i_m\leq T\text{ or }\eta^i_m\leq T\}}]\,.\end{aligned}$$

From Lemma 9.3.9 we obtain,

$$\mathbb{E}[|\hat{X}^{i,N,M}_T-\tilde{X}^{i,N,M}_T|^s\mathbb{1}_{\{\tau^i_m>T,\ \eta^i_m>T\}}]\leq C(m,s)h^s\,.$$

Also let us note,

$$\mathbb{E}[|E_r(T)^i|^2]\leq 2\mathbb{E}[|X^{i,N}_T|^2+|\tilde{X}^{i,N,M}_T|^2]\leq 2C\,,$$

where we have used Propositions 9.3.5 and 9.3.8. Hence for any $\epsilon>0$, we can choose δ such that,

$$\frac{\delta s}{2}\mathbb{E}[|E_r(T)^i|^2]\leq\frac{\epsilon}{3}\,.$$

By subadditivity of measures, $\mathbb{E}[\mathbb{1}_{\{\tau^i_m\leq T\text{ or }\eta^i_m\leq T\}}]\leq\mathbb{P}(\tau^i_m\leq T)+\mathbb{P}(\eta^i_m\leq T)$ and then Proposition 9.3.5, there exists m^* (dependent on δ), such that for $m\geq m^*$,

$$\frac{2-s}{2\delta^{s/(2-s)}}\mathbb{P}(\tau^i_m\leq T)\leq\frac{\epsilon}{3}\,.$$

Then, noting by Lemma 9.3.12,

$$\mathbb{E}[\sup_{0\leq t\leq T}|\hat{X}^{i,N,M}_{t\wedge\theta^i_m} - X^{i,N}_{t\wedge\theta^i_m}|^2] \leq C(m)h\,.$$

By Lemma 9.3.10, by taking h small enough for any $\tilde{\epsilon} > 0$, $\mathbb{P}(\eta^i_m < T) \leq \tilde{\epsilon}$. Hence, for any δ and m, we can take h small enough such that,

$$\begin{aligned}2^{s-1}\big(&\mathbb{E}[|X^{i,N}_T - \hat{X}^{i,N,M}_T|^s \mathbb{1}_{\{\tau^i_m>T,\ \eta^i_m>T\}}]\\ &+\mathbb{E}[|\hat{X}^{i,N,M}_T - \tilde{X}^{i,N,M}_T|^s \mathbb{1}_{\{\tau^i_m>T,\ \eta^i_m>T\}}]\big) + \frac{2-s}{2\delta^{s/(2-s)}}\mathbb{P}(\eta^i_m \leq T) \leq \frac{\epsilon}{3}\,,\end{aligned}$$

hence we can take $\epsilon \to 0$ by taking $h \to 0$, which completes the result. □

Chapter 10

Importance Sampling for McKean-Vlasov SDEs

The aim of this chapter is to develop efficient importance sampling algorithms for computing the expectations of functionals of solutions to MV-SDEs. We recall that the quantity of interest, is $\theta = \mathbb{E}[G(X)]$, which is approximated by the Monte Carlo estimator

$$\hat{\theta}^{N,M} = \frac{1}{N}\sum_{i=1}^{N} G(X^{i,N,M}).$$

The precision of this approximation is affected by three sources of error.

- The statistical error, that is the difference between $\hat{\theta}^{N,n}$ and $\mathbb{E}[G(X^{i,N,M})]$.
- The discretisation error, that is, the difference between $\mathbb{E}[G(X^{i,N,M})]$ and $\mathbb{E}[G(X^{i,N})]$.
- The propagation of chaos error of approximating the MV-SDE with the interacting particle system, that is, the difference between $\mathbb{E}[G(X^{i,N})]$ and $\mathbb{E}[G(X)]$.

The discretisation error of ordinary SDEs has been analysed by many authors, and it is well known that, e.g., under the Lipschitz assumptions the Euler scheme has weak convergence error of order $\frac{1}{M}$. It is of course well known, the standard deviation of the statistical error is of order of $\frac{1}{\sqrt{N}}$.

There has been some work detailing the error from the propagation of chaos as a function of N, essentially for G and X nice enough the weak error is also of order $\frac{1}{\sqrt{N}}$, see for example [KHO97] and [Bos04] for further details. In spite of this relatively slow convergence, many MV-SDEs have a reasonably "nice" dependence on the law which makes the particle approximation a good technique. On the other hand, one often wants to consider *rare events* in the context of the MV-SDE, and in this realm the statistical error will dominate the propagation of chaos error. The focus here is therefore on the statistical error of the Monte Carlo method. We will discuss the point of statistical against propagation of chaos error in more detail in Section 10.3.

Importance sampling is based on the following identity, valid for any probability measure $\mathbb{Q}$ (absolutely continuous with respect to $\mathbb{P}$)

$$\mathbb{E}[G(X)] = \mathbb{E}_{\mathbb{Q}}\left[\frac{d\mathbb{P}}{d\mathbb{Q}}G(X)\right].$$

The variance of the Monte Carlo estimator obtained by simulating X under the measure $\mathbb{Q}$ and correcting by the corresponding Radon-Nikodym density is different from that of the standard estimator, and can be made much smaller by a judicious choice of the sampling measure $\mathbb{Q}$.

Importance sampling is most effective in the context of *rare event simulation*, e.g., when the probability $\mathbb{P}[G(X) > 0]$ is small. Since the theory of large deviations is concerned with the study of probabilities of rare events, it is natural to use measure changes appearing in or inspired by the large deviations theory for importance sampling. We refer, e.g., to [DW04] and references therein for a review of this approach and to [GHS99], [GR08], [Rob10] for specific applications to financial models. The large deviations theory, on the one hand, simplifies the computation of the candidate importance sampling measure, and on the other hand, allows to define its optimality in a rigorous asymptotic framework.

Our Contribution. The main contribution of this work is two-fold. Firstly we show how one can apply a change of measure to MV-SDEs, and propose two algorithms that can carry this out: the *complete measure change* algorithm and the *decoupling* algorithm. In the complete measure change approach, the IS measure change is applied simultaneously in the coefficients and in the expectation to be evaluated. In the decoupling approach we first estimate the law of the solution in a first set of simulations without measure change and then perform a second set of simulations under the importance sampling measure using the approximate solution law computed in the first step.

Secondly, for both approaches, we use large deviations techniques to obtain an optimisation problem for the candidate measure change. We focus on the class of Cameron-Martin transforms, under which the measure change is given by

$$\frac{d\mathbb{Q}}{d\mathbb{P}}\Big|_{\mathcal{F}_T} = \mathcal{E}\Big(\int_0^T f_t \mathrm{d}W_t\Big) := \exp\Big(\int_0^T f_t \mathrm{d}W_t - \frac{1}{2}\int_0^T f_t^2 \mathrm{d}t\Big), \tag{10.0.1}$$

where f_t is a deterministic function. Following earlier works on the subject, we use the large deviations theory to construct a tractable proxy for the variance of $G(X)$ under the new measure. Of course, the presence of the interacting particle approximation introduces additional complexity at this point. Moreover, unlike the work of [GR08] which considered a very restrictive class of SDEs (the geometric Brownian motion), here we deal with a general class of MV-SDE where the drifts are of super-linear growth and satisfy a monotonicity type condition. This is very important in practice since many MV-SDEs fall into this category.

We then minimise the large deviations proxy to obtain a candidate optimal measure change for the two approaches that we consider. We find that the decoupling approach yields an easier optimisation problem than the complete measure change, which results in a high dimensional problem. However, by using exchangeability arguments the latter problem can be transformed into a far simpler two dimensional one. We implement both algorithms for two examples coming from the Kuramoto model from statistical physics and show that the variance of the importance sampling schemes is up to 3 orders of magnitude smaller than that of the standard Monte Carlo. Moreover, the computational only increases by a factor of 2–3 for the decoupling approach and is approximately the same as standard Monte Carlo for the complete measure change. We also estimate the propagation of chaos error and find that this is dominated by the statistical error by one order of magnitude. That being said, although the complete measure change appears to operate well in certain situations, it does rely on a change of measure which isn't too "large". We come back to this point throughout.

Concerning the measure change paradigm, in this work we focus on deterministic (open loop) measure changes over stochastic (feedback) measure changes. This is a decision one faces when using importance sampling and there are advantages and disadvantages to both. As pointed out in [GW97], deterministic measure changes may lead to detrimental results in terms of variance reduction, however, the increase in computational time of the importance sampling is overall negligible. Stochastic measure changes as discussed in [DW04] give improved variance reduction in far more generality, however, calculating the measure change is computationally burdensome, so the overall computational gain is less clear. As this is the first paper to marry importance sampling with MV-SDEs we feel it is beneficial to use deterministic based measure changes and leave stochastic measure changes as interesting future work. We provide precise conditions under which our deterministic measure change leads to an asymptotically optimal importance sampling estimator in the class of all possible measure changes. Further, one of our algorithms requires a measure changed propagation of chaos result to hold (Proposition 10.1.4) and it is not clear how to prove such a result if one uses stochastic measure changes.

The chapter is organised as follows. In Section 10.1 we discuss how importance sampling and measure changes can be carried out for MV-SDE, and in Section 10.2 we introduce our concept of optimality and identify the candidate optimal measure changes using the theory of large deviations. Section 10.3 illustrates numerically our results while proofs from Section 10.2

are carried out in Section 10.4.

10.1 Potential Importance sampling Algorithms

Leaving LDPs and the optimality of the IS (importance sampling) on the side, let us discuss how IS can be achieved for MV-SDEs with a given measure change.

Recall that MV-SDEs take the form (8.2.1). Because we change the measure we make explicit the dependence on the law of the solution process $\mu^X_{t,\mathbb{P}} = \mathbb{P} \circ X_t^{-1}$. If one knows the law μ^X beforehand , then one can treat the MV-SDE as a "standard" SDE and use IS as usual. However, typically one does not have access to the law, and the MV-SDE must be approximated by a so-called particle system approximation.

Remark 10.1.1 (Initial Condition). *Throughout this chapter we only consider MV-SDEs with deterministic initial conditions.*

The interacting particle system approximation. We approximate (8.2.1) (driven by the $\mathbb{P}$-Brownian motion $W^{\mathbb{P}}$), using an N-dimensional system of interacting particles. Let $i = 1, \dots, N$ and consider N particles $X^{i,N}$ satisfying the SDE with $X_0^{i,N} = x_0$

$$\mathrm{d}X_t^{i,N} = b\Big(t, X_t^{i,N}, \mu_t^{X,N}\Big)\mathrm{d}t + \sigma\Big(t, X_t^{i,N}, \mu_t^{X,N}\Big)\mathrm{d}W_t^{i,\mathbb{P}}, \quad \mu_t^{X,N}(\mathrm{d}x) := \frac{1}{N}\sum_{j=1}^{N} \delta_{X_t^{j,N}}(\mathrm{d}x) \tag{10.1.1}$$

where $\delta_{X_t^{j,N}}$ is the Dirac measure at point $X_t^{j,N}$, and the independent $\mathbb{P}$-Brownian motions $W^{i,\mathbb{P}}, i = 1, \dots, N$ (also independent of the BM $W^{\mathbb{P}}$ appearing in (8.2.1)). Due to the several changes of the measure throughout this section we keep track of which W we refer to.

Remark 10.1.2 (On the empirical measure $\mu_t^{X,N}$). *Unlike standard measures, empirical measures do not have dependence on the underlying measure $\mathbb{P}$, namely empirical measures are maps that depend on a sequence of $\omega^i \in \Omega$, thus one should write $\mu_t^{X,N}$ instead of $\mu_{t,\mathbb{P}}^{X,N}$. Of course, this is a pathwise statement, since the ω^i are generated under $\mathbb{P}$, the* distribution *of the empirical measure does depend on $\mathbb{P}$.*

Propagation of chaos. Again one can consider the system of non interacting particles

$$\mathrm{d}X_t^i = b(t, X_t^i, \mu_{t,\mathbb{P}}^{X^i})\mathrm{d}t + \sigma(t, X_t^i, \mu_{t,\mathbb{P}}^{X^i})\mathrm{d}W_t^{i,\mathbb{P}}, \quad X_0^i = x_0\,, \quad t \in [0,T]\,,$$

Then under nice enough conditions we obtain a pathwise propagation of result of (10.1.1) to this system (see Proposition 9.1.2).

Setup to change measures. When it comes to changing the measure under which we simulate we are also changing our approximation of the law. Since MV-SDEs depend explicitly on the law, this makes importance sampling more difficult. This will be one of the main points throughout this section.

Fix a deterministic square-integrable function $\dot{h} \in L_0^2(\mathbb{R})$. Then one can define the probability measure $\mathbb{Q}$ via the Girsanov transform $\frac{d\mathbb{Q}}{d\mathbb{P}}|_{\mathcal{F}_T} := \mathcal{E}(\int_0^T \dot{h}_t \mathrm{d}W_t^{\mathbb{P}})$, see (10.0.1), so that $\mathrm{d}W_t^{\mathbb{Q}} = \mathrm{d}W_t^{\mathbb{P}} - \dot{h}_t\mathrm{d}t$ is a $\mathbb{Q}$-Brownian motion. We note that the Radon-Nikodym density $\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}}|_{\mathcal{F}_t} = \mathcal{E}(\int_0^{\cdot} \dot{h}_s \mathrm{d}W_s^{\mathbb{P}})_t =: \mathcal{E}_t$ is itself the solution of the SDE

$$\mathrm{d}\mathcal{E}_t = \dot{h}_t\mathcal{E}_t\mathrm{d}W_t^{\mathbb{P}}, \quad \mathcal{E}_0 = 1 \quad \Rightarrow \quad \mathcal{E}_t = \exp\Big(\int_0^t \dot{h}_s \mathrm{d}W_s^{\mathbb{P}} - \frac{1}{2}\int_0^t |\dot{h}_s|^2\mathrm{d}s\Big).$$

Since $\mathbb{P}$ and $\mathbb{Q}$ are equivalent, one can also define $Z_t := \mathcal{E}_t^{-1} := \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}|_{\mathcal{F}_t}$. With our conditions on $\dot{h}$ it is also a straightforward task to show $\mathcal{E}_t$ and Z_t are in $\mathbb{S}^p$ for all $p \geq 1$.

Recall our goal: *estimate $\mathbb{E}_{\mathbb{P}}[G(X_T)] = \mathbb{E}_{\mathbb{Q}}[G(X_T)\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}]$ for some function G by simulating X under $\mathbb{Q}$*. In the following paragraphs we present two alternative ways to achieve this goal.

A running example. We present our algorithm in general setting with (8.2.1). For the sake of clarity and easiness of presentation, we often recourse to a particular class of MV-SDEs (under $\mathbb{P}$),

$$\mathrm{d}X_t = \hat{b}\big(t, X_t, \mathbb{E}_{\mathbb{P}}[f(X_t)]\big)\mathrm{d}t + \sigma \mathrm{d}W_t^{\mathbb{P}}, \quad X_0 = x_0\,, \quad t \in [0,T]\,. \tag{10.1.2}$$

with $\sigma \in \mathbb{R}_+$ and $f, \hat{b}$ nice*. We believe many of the arguments that are used at this level can be extended to cover more general MV-SDEs (such as higher order interactions). However, obtaining analogous results to those of standard MV-SDEs, such as propagation of chaos, is made more challenging by the inclusion of the measure change. Therefore, these have to be considered on a case by case basis.

10.1.1 Fixing the Empirical Law - a decoupling argument

An obvious way to solve the problem of IS is to approximate the law of the MV-SDE under $\mathbb{P}$ and use that as a fixed input to a new equation which will be simulated under $\mathbb{Q}$. In this set up the McKean-Vlasov SDE turns into an SDE with random coefficients. The algorithm is as follows.

1. Use (10.1.1) with N particles to approximate (8.2.1). Use some numerical scheme (under $\mathbb{P}$, say Euler) to simulate the particles in time, calculating an empirical law over $[0,T]$. This gives an approximation for the empirical law μ_t^N which is then fixed.

   Define a new SDE, approximating the original MV-SDE (8.2.1), which is now a *standard SDE with random coefficients*

   $$\mathrm{d}\bar{X}_t = b(t, \bar{X}_t, \mu_t^N)\mathrm{d}t + \sigma(t, \bar{X}_t, \mu_t^N)\mathrm{d}W_t^{\mathbb{P}}, \quad \bar{X}_0 = x_0, \tag{10.1.3}$$

   where $W^{\mathbb{P}}$ is a $\mathbb{P}$-Brownian motion independent of the $\{W^{i,\mathbb{P}}\}_{i=1,\cdots,N}$ appearing in (10.1.1). SDEs with random coefficients appear typically in optimal control, hence the reader can consult texts such as [YZ99, Chapter 1] for further details on existence uniqueness of such SDEs.

2. Change the probability measure to $\mathbb{Q}$, which is our importance sampling measure change. Simulate (10.1.3) under this new measure, i.e.

   $$\mathrm{d}\bar{X}_t = \Big(b(t, \bar{X}_t, \mu_t^N) + \dot{h}_t \sigma(t, \bar{X}_t, \mu_t^N)\Big)\,\mathrm{d}t + \sigma(t, \bar{X}_t, \mu_t^N)\mathrm{d}W_t^{\mathbb{Q}}, \quad \bar{X}_0 = x_0\,.$$

3. This second run is therefore standard importance sampling, but the SDE has random coefficients i.e. the empirical law is random.

We will refer to algorithms of this form as *Decoupling Algorithms*. This scheme has the disadvantage in that it requires twice the amount of simulation and one will require a handle on the error coming from the original approximation of the law.

It is not a requirement to use interacting particles to approximate the law of the SDE, any approximation will work. The goal here is to make the SDEs independent.

10.1.2 Complete Measure Change

An alternative is to change the measure under which we are simulating in the coefficients *and* the Brownian motion. This is not a simple problem and as far as we are aware changing the measure of a MV-SDE and its particle approximation is not discussed elsewhere in the literature (for this purpose†), we therefore provide a discussion along with the pitfalls here. This is more complex than the decoupled case and for clarity we use (10.1.2) throughout.

*We use $\hat{b}$ here since it takes the expectation rather than a measure input.

†Measures changes for MV-SDE appear in methods requiring to remove the drift altogether, for instance in establishing weak solutions to MV-SDEs, see e.g. [DG87].

The measure changed version of (10.1.2) takes the form,

$$\begin{aligned}\mathrm{d}X_t &= \Big(\hat{b}(t, X_t, \mathbb{E}_{\mathbb{P}}[f(X_t)]) + \sigma\dot{h}_t\Big)\mathrm{d}t + \sigma\mathrm{d}W_t^{\mathbb{Q}}\\ &= \Big(\hat{b}(t, X_t, \mathbb{E}_{\mathbb{Q}}\Big[f(X_t)Z_t\Big]) + \sigma\dot{h}_t\Big)\mathrm{d}t + \sigma\mathrm{d}W_t^{\mathbb{Q}}.\end{aligned}$$

where again $Z := \mathcal{E}^{-1}$.

In view of simulation, we re-write the measure changed MV-SDE as a system

$$\mathrm{d}X_t = \Big(\hat{b}\Big(t, X_t, \mathbb{E}_{\mathbb{Q}}\big[F(X_t, Z_t)\big]\Big) + \sigma\dot{h}_t\Big)\mathrm{d}t + \sigma\mathrm{d}W_t^{\mathbb{Q}}, \quad \text{and} \quad \mathrm{d}Z_t = \dot{h}_t Z_t \mathrm{d}W_t^{\mathbb{Q}}, \quad Z_0 = 1\,,$$

where $F(x, y) = f(x)y$. We now write the interacting particle system for the pair X, Z under $\mathbb{Q}$:

$$\begin{aligned}\mathrm{d}X_t^{i,N} &= \Big(\hat{b}\big(t, X_t^{i,N}, \frac{1}{N}\sum_{j=1}^{N} F(X_t^{j,N}, Z_t^{j,N})\big) + \sigma\dot{h}_t\Big)\mathrm{d}t + \sigma\mathrm{d}W_t^{i,\mathbb{Q}}, \qquad (10.1.4)\\ \mathrm{d}Z_t^{i,N} &= \dot{h}_t Z^{i,N}\mathrm{d}W_t^{i,\mathbb{Q}}, \quad Z_0^{i,N} = 1\,.\end{aligned}$$

The importance sampling estimator of $\theta = \mathbb{E}^{\mathbb{P}}[G(X_T)]$ then takes the form

$$\hat{\theta}_h = \frac{1}{N}\sum_{i=1}^{N} Z_T^{i,N} G(X_T^{i,N}). \qquad (10.1.5)$$

Remark 10.1.3. *One may be tempted to use a different approach, namely first apply an interacting particle approximation under $\mathbb{P}$ which yields*

$$\mathrm{d}X_t^{i,N} = \hat{b}\big(t, X_t^{i,N}, \frac{1}{N}\sum_{j=1}^{N} f(X_t^{j,N})\big)\mathrm{d}t + \sigma\mathrm{d}W_t^{i,\mathbb{P}},$$

and then change the measure for the particle system, writing

$$\mathrm{d}X_t^{i,N} = \Big(\hat{b}\big(t, X_t^{i,N}, \frac{1}{N}\sum_{j=1}^{N} f(X_t^{j,N})\big) + \sigma\dot{h}_t\Big)\mathrm{d}t + \sigma\mathrm{d}W_t^{i,\mathbb{Q}},$$

where we have taken the same $\dot{h}$ for every Brownian motion in order for all particles to have the same law. However, it is easy to see by the standard propagation of chaos result that as $N \to \infty$, this particle system converges to the solution of the MV-SDE (see Proposition 9.1.2)

$$\mathrm{d}X_t = \Big(\hat{b}\big(t, X_t, \mathbb{E}^{\mathbb{Q}}[X_t]\big) + \sigma\dot{h}_t\Big)\mathrm{d}t + \sigma\mathrm{d}W_t^{\mathbb{Q}} = \hat{b}\big(t, X_t, \mathbb{E}^{\mathbb{Q}}[X_t]\big)\mathrm{d}t + \sigma\mathrm{d}W_t^{\mathbb{P}},$$

which is not what one is looking for.

To state a propagation of chaos result for the particle system (10.1.4) we introduce the auxiliary system of non-interacting particles,

$$\begin{aligned}\mathrm{d}X_t^{i} &= \Big(\hat{b}\Big(t, X_t^{i}, \mathbb{E}_{\mathbb{Q}}\big[F(X_t^{i}, Z_t^{i})\big]\Big) + \sigma\dot{h}_t\Big)\mathrm{d}t + \sigma\mathrm{d}W_t^{i,\mathbb{Q}}\,, \qquad (10.1.6)\\ \mathrm{d}Z^{i} &= \dot{h}_t Z^{i}\mathrm{d}W_t^{i,\mathbb{Q}}, \quad Z^{i} = 1\,.\end{aligned}$$

Proposition 10.1.4. *Consider the following measure changed MV-SDE (see* (10.1.6)*),*

$$\mathrm{d}X_t^{i} = \Big(\hat{b}\Big(t, X_t^{i}, \mathbb{E}_{\mathbb{Q}}\Big[f(X_t^{i})Z_t^{i}\Big]\Big) + \sigma\dot{h}_t\Big)\mathrm{d}t + \sigma\mathrm{d}W_t^{i,\mathbb{Q}}\,, \quad \mathrm{d}Z_t^{i} = \dot{h}_t Z_t^{i}\mathrm{d}W_t^{i,\mathbb{Q}}, \quad Z_0^{i} = 1\,, \qquad (10.1.7)$$

where $\hat{b}$ is continuous in time, $\hat{b}$ and f are Lipschitz in space, and $\hat{b}$ is a bounded Lipschitz function in its third variable. Let $X_t^{i,N}$, denote the corresponding particle approximation (see (10.1.4)*). Then* *the following pathwise propagation of chaos result holds,*

$$\lim_{N\to\infty} \sup_{1\leq i\leq N} \mathbb{E}_{\mathbb{Q}}\left[\sup_{0\leq t\leq T} |X_t^{i,N} - X_t^i|^2\right] = 0\,.$$

This proposition may be used to analyze the convergence of the Monte Carlo estimator (10.1.5). Indeed, due to the fact that there is no coupling (or law dependency) in $Z_t^{i,N}$, $Z^{i,N} = Z^i$ and $\hat{\theta}_h$ can be represented as follows.

$$\hat{\theta}_h = \frac{1}{N}\sum_{i=1}^{N} Z_T^i G(X_T^i) + \frac{1}{N}\sum_{i=1}^{N} Z_T^i (G(X_T^{i,N}) - G(X_T^i)).$$

The first term above converges to θ as $N \to \infty$ by the law of large numbers, and the second term can be shown, e.g., to converge to zero in probability using Proposition 10.1.4 if G is sufficiently regular.

Proof of Proposition 10.1.4. The idea of the proof is to appeal to a Grönwall type inequality, but this is made difficult due to the presence of Z term in (10.1.7). One can note, due to the assumptions on the coefficients of the SDE, all p-moments exist. Using our prescribed form of the MV-SDE we obtain,

$$|X_t^{i,N} - X_t^i|^2 \leq C\int_0^t \Big|\hat{b}\Big(s, X_s^{i,N}, \frac{1}{N}\sum_{j=1}^{N} f(X_s^{j,N})Z_s^j\Big) - \hat{b}\left(s, X_s^i, \mathbb{E}_{\mathbb{Q}}[f(X_s^i)Z_s^i]\right)\Big|^2 \mathrm{d}s\,.$$

Let $s \in [0, T]$, then introduce the terms, $\hat{b}\Big(s, X_s^i, \frac{1}{N}\sum_{j=1}^{N} f(X_s^{j,N})Z_s^j\Big)$ and $\hat{b}\Big(s, X_s^i, \frac{1}{N}\sum_{j=1}^{N} f(X_s^j)Z_s^j\Big)$, where the empirical measure in the second term is the one constructed from the i.i.d. SDEs in (10.1.7), hence each X^j corresponds to a independent realisation of the MV-SDE, namely it has the correct distribution. Splitting the original difference into three, we use the Lipshitz property in space for the first one, to obtain,

$$\Big|\hat{b}\Big(s, X_s^{i,N}, \frac{1}{N}\sum_{j=1}^{N} f(X_s^{j,N})Z_s^j\Big) - \hat{b}\Big(s, X_s^i, \frac{1}{N}\sum_{j=1}^{N} f(X_s^{j,N})Z_s^j\Big)\Big|^2 \leq C|X_s^{i,N} - X_s^i|^2\,.$$

For the second difference we use the fact that $\hat{b}$ is bounded along with the Lipschitz property in the third variable, which yields

$$\begin{aligned}
&\Big|\hat{b}\Big(s, X_s^i, \frac{1}{N}\sum_{j=1}^{N} f(X_s^{j,N})Z_s^j\Big) - \hat{b}\Big(s, X_s^i, \frac{1}{N}\sum_{j=1}^{N} f(X_s^j)Z_s^j\Big)\Big|^2 \\
&\leq C\Big|\hat{b}\Big(s, X_s^i, \frac{1}{N}\sum_{j=1}^{N} f(X_s^{j,N})Z_s^j\Big) - \hat{b}\Big(s, X_s^i, \frac{1}{N}\sum_{j=1}^{N} f(X_s^j)Z_s^j\Big)\Big| \leq C\frac{1}{N}\sum_{j=1}^{N} Z_s^j|X_s^{j,N} - X_s^j|\,.
\end{aligned}$$

Finally, again from the Lipschitz property we obtain,

$$\Big|\hat{b}\Big(s, X_s^i, \frac{1}{N}\sum_{j=1}^{N} f(X_s^j)Z_s^j\Big) - \hat{b}\left(s, X_s^i, \mathbb{E}_{\mathbb{Q}}[f(X_s^i)Z_s^i]\right)\Big|^2 \leq C\Big|\frac{1}{N}\sum_{j=1}^{N} f(X_s^j)Z_s^j - \mathbb{E}_{\mathbb{Q}}[f(X_s^i)Z_s^i]\Big|^2\,.$$

Hence the following bound holds,

$$\begin{aligned}
\mathbb{E}_{\mathbb{Q}}\Big[\sup_{0\leq t\leq T} |X_t^{i,N} - X_t^i|^2\Big] \leq C\int_0^T &\mathbb{E}_{\mathbb{Q}}[|X_s^{i,N} - X_s^i|^2] + \frac{1}{N}\sum_{j=1}^{N} \mathbb{E}_{\mathbb{Q}}\Big[Z_s^j|X_s^{j,N} - X_s^j|\Big] \\
&+ \mathbb{E}_{\mathbb{Q}}\Big[\Big|\frac{1}{N}\sum_{j=1}^{N} f(X_s^j)Z_s^j - \mathbb{E}_{\mathbb{Q}}[f(X_s^i)Z_s^i]\Big|^2\Big]\mathrm{d}s\,.
\end{aligned}$$

One can use Cauchy-Schwarz along with the properties of Z to obtain,

$$\mathbb{E}_{\mathbb{Q}}\Big[Z_s^j|X_s^{j,N}-X_s^j|\Big]\le C\mathbb{E}_{\mathbb{Q}}\Big[|X_s^{j,N}-X_s^j|^2\Big]^{\frac{1}{2}}\le C\mathbb{E}_{\mathbb{Q}}\Big[\sup_{0\le u\le s}|X_u^{j,N}-X_u^j|^2\Big]^{\frac{1}{2}}.$$

Although at first it appears one cannot use Grönwall here, there is a nonlinear generalisation due to Perov (see [MPF12, Theorem 1, p360]) which we can use since the nonlinear term on the RHS is square root of the term on the left. Finally, take the supremum over i and using the fact that the variables $f(X_s^j)Z_s^j$ are i.i.d. and square integrable, we obtain,

$$\begin{aligned}\sup_{1\le i\le N}\mathbb{E}_{\mathbb{Q}}\Big[\sup_{0\le t\le T}|X_t^{i,N}-X_t^i|^2\Big]&\le Ce^C\int_0^T\mathbb{E}_{\mathbb{Q}}\Big[\big|\frac{1}{N}\sum_{j=1}^N f(X_s^j)Z_s^j-\mathbb{E}_{\mathbb{Q}}[f(X_s^i)Z_s^i]\big|^2\Big]\mathrm{d}s\\&\le\frac{Ce^C}{N}\int_0^T\mathbb{E}_{\mathbb{Q}}\Big[\big|f(X_s^1)Z_s^1-\mathbb{E}_{\mathbb{Q}}[f(X_s^1)Z_s^1]\big|^2\Big]\mathrm{d}s\to 0\end{aligned}$$

as $N\to\infty$, which concludes the proof. □

The Complete Measure Change Algorithm We now describe the algorithm for simulating a general MV-SDE under a complete measure change.

1. Simulate the following particle system for the MV-SDE after the measure change:

$$\begin{aligned}\mathrm{d}X_t^{i,N}&=\Big(b\Big(t,X_t^{i,N},\frac{1}{N}\sum_{j=1}^N Z_t^j\delta_{X_t^{j,N}}\Big)+\dot{h}_t\sigma\Big(t,X_t^{i,N},\frac{1}{N}\sum_{j=1}^N Z_t^j\delta_{X_t^{j,N}}\Big)\Big)\mathrm{d}t\\&\quad+\sigma\Big(t,X_t^{i,N},\frac{1}{N}\sum_{j=1}^N Z_t^j\delta_{X_t^{j,N}}\Big)\mathrm{d}W_t^{i,\mathbb{Q}},\\\mathrm{d}Z_t^i&=\dot{h}_tZ_t^i\mathrm{d}W_t^{i,\mathbb{Q}},\quad Z_0^i=1.\end{aligned}$$

2. Compute the importance sampling estimator using the following formula:

$$\hat{\theta}_h=\frac{1}{N}\sum_{i=1}^N Z_T^{i,N}G(X_T^{i,N}).$$

We will refer to algorithms of this form as *Complete Measure Change Algorithms*. An advantage one can immediately see is that one simulates the particles only once. A key disadvantage is that the importance sampling to estimate the object of interest $\mathbb{E}[G(X_T)]$, may yield a poorer estimation of the original law μ and the term $\mathbb{E}_{\mathbb{Q}}[f(X_t)Z_t]$ in (10.1.6). We will discuss this in Section 10.3.

10.2 Optimal Importance Sampling for McKean-Vlasov SDEs

The previous section detailed algorithms for simulating MV-SDEs under an arbitrary change of measure. We now want to use the theory of large deviations to determine, in a certain optimal way, a measure change which will reduce the variance of the estimate.

An important point here is that we will be using the LDP for Brownian motion, rather than the MV-SDEs. There are several works dealing with Large Deviations for MV-SDEs and their associated interacting particles systems, see [BDF12], [Fis14], [dRST17] but such results are not of use to us here since we must be able to (cheaply) simulate the SDE after the change of measure. We therefore restrict to the Girsanov measure change since we know how the SDE changes under the measure change.

In this section we first show how the LDP framework can be applied to both algorithms to yield a simplified optimisation problem for finging the asymptotically optimal measure change

(Theorems 10.2.9 and 10.2.7) and then demonstrate how these simplified optimization problems may be solved in practice.

10.2.1 Preliminaries

We recall some of the main concepts for importance sampling with LDP, see [GR08] for further discussion. We denote by $\mathbb{W}^d_T$ the standard d-dimensional Wiener space of continuous functions over the time interval $[0,T]$ which are zero at time zero and in the one-dimensional case we simply write $\mathbb{W}_T$ instead of $\mathbb{W}^1_T$. This space is endowed with the topology of uniform convergence and with the usual Wiener measure $\mathbb{P}$, defined on the completed filtration $\mathcal{F}_T$, which makes the process $\mathbf{W}_t(x) = x_t$ with $x \in \mathbb{W}^d_T$ a standard d-dimensional Brownian motion.

The goal is to estimate the expected value of some functional $\tilde{G} : \mathbb{W}^d_T \to \mathbb{R}_+$ continuous in the uniform topology ($\tilde{G}$ is explained later). For the change of measure, one considers a Girsanov transform where the allowed functions are from the Cameron-Martin space, i.e. (if $d = 1$ we just write $\mathbb{H}_T = \mathbb{H}^1_T$)

$$\mathbb{H}^d_T = \left\{ h : [0,T] \mapsto \mathbb{R}^d : \ h_0 = 0\,,\ h_\cdot = \int_0^\cdot \dot{h}_t dt\,,\ \int_0^T |\dot{h}_t|^2 \,\mathrm{d}t < \infty \ \text{ i.e. } \ \dot{h}_t \in L^2_t(\mathbb{R}^d) \right\}.$$

For any deterministic drift $h \in \mathbb{H}^d_T$, the stochastic exponential defines the Radon-Nikodym derivative for an equivalent measure $\mathbb{Q}$ namely, ($W^{\mathbb{P}}$ is a standard $\mathbb{P}$-Brownian motion)

$$\frac{\mathrm{d}\mathbb{Q}}{\mathrm{d}\mathbb{P}} = \exp\Big(\int_0^T \dot{h}_t \mathrm{d}W^{\mathbb{P}}_t - \frac{1}{2}\int_0^T (\dot{h}_t)^2 \,\mathrm{d}t \Big). \tag{10.2.1}$$

Under this new measure $\mathbb{Q}$, the process $W^{\mathbb{Q}}_\cdot = W^{\mathbb{P}}_\cdot - h_\cdot$ is a standard d-dimensional $\mathbb{Q}$-Brownian motion.

Standing assumptions We consider MV-SDEs with nonlinear interaction between the SDE and its law. In this section we concentrate on one-dimensional SDEs of the form,

$$\mathrm{d}X_t = b(t, X_t, \mu_t)\mathrm{d}t + \sigma \mathrm{d}W_t, \qquad X_0 = x_0. \tag{10.2.2}$$

Throughout this section we will refer to the following assumptions (similar to assumptions in Section 8.2), for functions $b : [0,T] \times \mathbb{R} \times \mathcal{P}_2(\mathbb{R}) \to \mathbb{R}$ and $\sigma > 0$ constant.

Assumption 10.2.1. *Assume that b is Lipschitz in the sense that $\exists L > 0$ such that $\forall t \in [0,T]$, $\forall x, x' \in \mathbb{R}$ and $\forall \mu, \mu' \in \mathcal{P}_2(\mathbb{R})$ we have that*

$$|b(t,x,\mu) - b(t,x',\mu')| \leq L(|x - x'| + W^{(2)}(\mu, \mu')).$$

Moreover, $\forall\ x \in \mathbb{R}$ and $\mu \in \mathcal{P}_2(\mathbb{R})$, b is continuous over the interval $[0,T]$.

Assumption 10.2.2. *Assume b satisfies the monotone growth and local Lipschitz conditions in Assumption 8.2.1. Further, $\forall\ x \in \mathbb{R}$ and $\mu \in \mathcal{P}_2(\mathbb{R})$, let b be continuous in time over the interval $[0,T]$.*

In view of Section 8.2, either of these assumptions yield the existence of a unique strong solution to (10.2.2). We further use the following assumption for the terminal function G. Note that this assumption is on G as a function of the SDE, rather than the driving Brownian motion as is the case in [GR08].

Assumption 10.2.3. *The functional G is non-negative, continuous and satisfies the following growth condition*

$$\log(G(x)) \leq C_1 + C_2 \sup_{t \in [0,T]} |x_t|^\alpha\,,$$

for $x : [0,T] \mapsto \mathbb{R}$ a continuous function starting at x_0 where $C_1,\ C_2$ are positive constants and $\alpha \in [1,2)$.

The notion of "optimality" for the measure change we use is so-called *asymptotically optimal*, as defined in[‡] [GHS99]. Following the approach of [GHS99], we want to estimate $\mathbb{E}[\exp(\log(G(X)))]$. Here we perform a measure change for the Brownian motion, so for ease of writing let us define $F(W) := \log(G(X(W)))$ and consider the more general problem of estimating,

$$\alpha(\epsilon) := \mathbb{E}[\exp(F(\sqrt{\epsilon}W)/\epsilon)], \qquad \text{for } \epsilon > 0.$$

This is our original problem when $\epsilon = 1$, and we can use Varadhan's lemma to understand this quantity as $\epsilon \to 0$, this is referred to as *small noise asymptotics*. We now consider a general estimator for this quantity $\hat{\alpha}(\epsilon)$ (there is no requirement for $\hat{\alpha}$ to be based on a deterministic measure change). At this point we have no conditions on these estimators so we follow definition [GHS99, Definition 2.1].

Definition 10.2.4. *A family of estimators* $\{\hat{\alpha}(\epsilon)\}$ *is said to be* asymptotically relatively unbiased *if the following holds,*

$$\frac{\mathbb{E}[\hat{\alpha}(\epsilon)] - \alpha(\epsilon)}{\alpha(\epsilon)} \to 0 \quad \textit{as } \epsilon \to 0\,.$$

The above definition yields estimators that in some sense converge, but we are interested in comparing such estimators and for this we look at their second moment.

Definition 10.2.5. *A family of relatively unbiased estimators* $\{\hat{\alpha}_0(\epsilon)\}$ *is said to be* asymptotically optimal *if,*

$$\limsup_{\epsilon\to 0} \epsilon \log \mathbb{E}[\hat{\alpha}_0(\epsilon)^2] = \inf_{\{\hat{\alpha}(\epsilon)\}} \limsup_{\epsilon\to 0} \epsilon \log \mathbb{E}[\hat{\alpha}(\epsilon)^2],$$

where the infimum is over all asymptotically relatively unbiased estimators.

One of the goals of this section will be obtaining conditions when measure changes of type (10.2.1) are asymptotically optimal. As it turns out, using this definition it is not difficult to obtain a necessary and sufficient condition for asymptotic optimality, a similar argument is given in [GHS99, pg 126]. Let us consider some asymptotic unbiased estimator $\hat{\alpha}$, and define the difference $\Delta(\epsilon) := \mathbb{E}[\hat{\alpha}(\epsilon)] - \alpha(\epsilon)$, it is a straightforward consequence of Jensen's inequality and some rearranging,

$$\log(\mathbb{E}[\hat{\alpha}(\epsilon)^2]) \geq 2\log(\mathbb{E}[\hat{\alpha}(\epsilon)]) = 2\log(\mathbb{E}[\alpha(\epsilon)]) + O(\Delta(\epsilon)/\alpha(\epsilon)) \xrightarrow{\epsilon\to 0} 2\log(\mathbb{E}[\alpha(\epsilon)])\,.$$

Thus we have a lower bound for an estimator, moreover, note that this implies the degenerate estimator $\hat{\alpha}(\epsilon) = \alpha(\epsilon)$ is asymptotically optimal, since α is not random. One can use Varadhan's lemma and Schilder's theorem (see Section 10.4.1) since we are dealing with Brownian motion to obtain,

$$\limsup_{\epsilon\to 0} 2\epsilon \log(\alpha(\epsilon)) = \sup_{u\in\mathbb{H}_T^d} \left\{ 2\log(G(X(u))) - \int_0^T |\dot{u}_t|^2 \mathrm{d}t \right\}. \tag{10.2.3}$$

Therefore any estimator which equals the RHS of this expression is asymptotically optimal. Depending on which algorithm we use this will be a slightly different expression but the argument to obtain the bound is the same.

10.2.2 The decoupling algorithm

We first consider the decoupling algorithm presented in Section 10.1.1. We build μ_t^N, from an independent N-particle system which is simulated under a numerical scheme, and then consider

[‡] A related but slightly weaker definition of optimality is used in [GR08].

the following approximation of SDE[§] (10.2.2),

$$\mathrm{d}\overline{X}_t = b(t, \overline{X}_t, \mu^N_t)\mathrm{d}t + \sigma \mathrm{d}W_t, \qquad X_0 = x_0\,. \tag{10.2.4}$$

In order to distinguish the current SDE from the previous particle approximation we introduce a so-called copy space (see for example [BLP+17]) $(\tilde{\Omega}, \tilde{\mathcal{F}}, (\tilde{\mathcal{F}}_t)_{t\geq 0}, \tilde{\mathbb{P}})$ (with the usual conditions and $\tilde{\mathcal{F}}_t$ is the augmented filtration over the N-dimensional Brownian motion). The N-system SDEs used to approximate this measure is then defined on this space, hence (10.2.4) is defined on the product space $(\Omega, \mathcal{F}, \mathbb{P}) \otimes (\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{\mathbb{P}})$.

Our aim is now to minimize over $h \in \mathbb{H}_T$ the variance conditional on the trajectory of μ^N:

$$\mathbb{E}_{\mathbb{P}\otimes\tilde{\mathbb{P}}}\big[\,G(\overline{X}_T)^2\mathcal{E}_T^{-1}\big|\tilde{\mathcal{F}}_T\big], \quad \mathrm{d}\mathcal{E}_t = \dot{h}_t\mathcal{E}_t\mathrm{d}W^{\mathbb{P}}_t, \quad \mathcal{E}_0 = 1,$$

and we make use of small noise asymptotics in order to write this variance in a "LDP" tractable form, hence we define, for $h \in \mathbb{H}_T$

$$L(h;\mu^N) := \limsup_{\epsilon\to 0} \epsilon \log \mathbb{E}_{\mathbb{P}\otimes\tilde{\mathbb{P}}}\left[\exp\left(\frac{1}{\epsilon}\Big(2\log(\overline{G}(\sqrt{\epsilon}W)) - \int_0^T \sqrt{\epsilon}\dot{h}_t\mathrm{d}W_t + \frac{1}{2}\int_0^T \dot{h}_t^2\mathrm{d}t\Big)\right)\Big|\tilde{\mathcal{F}}_T\right], \tag{10.2.5}$$

where $\overline{G}(W) := G(\overline{X}(W))$. One should also keep in mind that $\overline{G}$ also depends on μ^N, however, we suppress this notation for ease of presentation.

Remark 10.2.6. *In* (10.2.5)*, we have a conditional expectation, thus $L(h;\mu^N)$ is technically a random variable in $\tilde{\Omega}$. This is not typically the case when using Varadhan's lemma, however, because the random variable is independent of the Brownian motion and $\overline{G}$ is still $\tilde{\mathbb{P}}$-a.s. continuous w.r.t. the Brownian motion (Section 10.4.2), upon checking the moment condition, we are still able to use Varadhan's lemma, $\tilde{\mathbb{P}}$-a.s..*

Theorem 10.2.7. *Let Assumptions 10.2.3 and 10.2.2 hold and fix $\tilde{\omega} \in \tilde{\Omega}$ (and thus μ^N). Furthermore assume that there exists $u \in \mathbb{H}_T$ such that $\overline{G}(u) > 0$. Then the following statements hold:*

i. Let $h \in \mathbb{H}_T$ such that $\dot{h}$ is of finite variation. Then Varadhan's lemma holds for the small noise asymptotics, namely we can rewrite (10.2.5) *as,*

$$L(h;\mu^N) = \sup_{u\in\mathbb{H}_T}\left\{2\log(\overline{G}(u)) - \int_0^T \dot{h}_t\dot{u}_t\mathrm{d}t + \frac{1}{2}\int_0^T \dot{h}_t^2\mathrm{d}t - \frac{1}{2}\int_0^T \dot{u}_t^2\mathrm{d}t\right\} \quad \tilde{\mathbb{P}}\text{-a.s.}\,. \tag{10.2.6}$$

ii. There exists an $h^* \in \mathbb{H}_T$ which minimizes (10.2.6)*.*

iii. Consider a simplified optimization problem

$$\sup_{u\in\mathbb{H}_T}\left\{2\log(\overline{G}(u)) - \int_0^T \dot{u}_t^2\mathrm{d}t\right\}. \tag{10.2.7}$$

There exists a maximizer h^{**} for this problem. If

$$L(h^{**};\mu^N) = 2\log(\overline{G}(h^{**})) - \int_0^T (\dot{h}^{**}_t)^2\mathrm{d}t\,, \tag{10.2.8}$$

then h^{**} defines an asymptotically optimal measure change and is the unique maximizer of (10.2.7)*.*

All of these results are $\tilde{\mathbb{P}}$-a.s. since the particle system yields a random measure from $\tilde{\Omega}$. The proof of this theorem requires several auxiliary results which we defer to Section 10.4.2. One should also note that the requirement for $\overline{G} > 0$ for some u is not restrictive, it is purely there

[§] The measure, μ^N is a random measure, but is independent of the process $\overline{X}$ thus we have decoupled the SDE.

for technical reasons since one cannot have a maximiser if $\log(\overline{G}(u)) = -\infty$ for all $u \in \mathbb{H}_T$. The assumption that $\dot{h}$ has finite variation is necessary to establish the continuity of the functional in Varadhan's lemma.

Remark 10.2.8 (Concavity of $\log(\overline{G})$ and asymptotic optimality)**.** *Consider the problem of minimizing* (10.2.6) *and assume that one can interchange the inf and the sup. Then,*

$$\begin{aligned}\inf_{h\in\mathbb{H}_T} L(h;\mu^N) &= \sup_{u\in\mathbb{H}_T}\inf_{h\in\mathbb{H}_T}\left\{2\log(\overline{G}(u)) - \int_0^T \dot{h}_t\dot{u}_t\mathrm{d}t + \frac{1}{2}\int_0^T \dot{h}_t^2\mathrm{d}t - \frac{1}{2}\int_0^T \dot{u}_t^2\mathrm{d}t\right\}\\ &= \sup_{u\in\mathbb{H}_T}\left\{2\log(\overline{G}(u)) - \int_0^T \dot{u}_t^2\mathrm{d}t\right\}\end{aligned}$$

because the inner problem is solved by $h = u$. Therefore, a sufficient condition for an asymptotically optimal measure change of type (10.2.1) *is the exchangeability of inf and sup above. Since L is a convex function in h, and the integral terms in* (10.2.6) *are concave in u, a sufficient condition for such exchangeability is that $\log(\overline{G})$ is concave. Indeed, in the case of convex-concave functions we can invoke the minimax principle to swap infimum and supremum, see* [ET99, pg. 175] *for example.*

In [GR08]*, the process X was a geometric Brownian Motion and the authors were able to explicitly link the concavity of $\log(\overline{G})$ with the properties of the function G. Here the dependence of $\overline{G}$ on the Brownian motion is more complex, and it appears to be difficult to check concavity. Hence, in general one has to check numerically whether* (10.2.8) *holds. However, even if* (10.2.8) *fails, one can still use h^{**} to construct a candidate importance sampling measure if this is justified by superior numerical performance.*

10.2.3 The complete measure change algorithm

Here we focus on the algorithm discussed in Section 10.1.2. Recall that we are interested in evaluating, $\mathbb{E}_\mathbb{P}[G(X)]$. We now change the measure to $\mathbb{Q}$ and calculate the variance,

$$\text{Var}_\mathbb{Q}\Big[G(X)\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}\Big] = \mathbb{E}_\mathbb{P}\Big[G(X)^2\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}\Big] - \mathbb{E}_\mathbb{P}\Big[G(X)\Big]^2.$$

Minimising the variance is equivalent to minimize the first term in the RHS. As a first step to constructing a tractable proxy for this variance we consider a particle approximation of X:

$$\mathrm{d}X_t^{i,N} = b\left(t, X_t^{i,N}, \frac{1}{N}\sum_{j=1}^N \delta_{X_t^{j,N}}\right)\mathrm{d}t + \sigma\mathrm{d}W_t^{i,\mathbb{P}}, \quad X_0^{i,N} = x_0, \tag{10.2.9}$$

$$\mathrm{d}\mathcal{E}_t^i = \dot{h}_t\mathcal{E}_t^i\mathrm{d}W_t^{i,\mathbb{P}}, \quad \mathcal{E}_0^i = 1, \tag{10.2.10}$$

where $W^{i,\mathbb{P}}$ denotes the driving $\mathbb{P}$-Brownian motion of particle i, and all $W^{i,\mathbb{P}}$s are independent of each other. We approximate $\mathbb{E}_\mathbb{P}[G^2(X)(\mathcal{E}_T)^{-1}]$ with $\mathbb{E}_\mathbb{P}[G^2(X^{i,N})(\mathcal{E}_T^{i,N})^{-1}]$. Since $\mathcal{E}^i = \mathcal{E}^{i,N}$ (due to the absence of cross dependency), one can equivalently minimize

$$\mathbb{E}_\mathbb{P}\big[\ G^2(X^{i,N})(\mathcal{E}_T^i)^{-1}\ \big], \ \text{over all } h \in \mathbb{H}_T. \tag{10.2.11}$$

In order to use the LDP theory to minimize (10.2.11), we define $\tilde{G}$ as the functional dependent on the underlying $\mathbb{P}$-Brownian motions, i.e., for all $i \in \{1,\dots,N\}$, $\tilde{G}_i : \mathbb{W}_T^N \mapsto \mathbb{R}$, where, $\tilde{G}_i(W^1,\dots,W^N) := G(X^{i,N}(W^1,\dots,W^N))$. The corresponding small noise asymptotics takes the following form:

$$\bar{L}(h) := \limsup_{\epsilon\to 0} \epsilon \log \Bigg(\mathbb{E}_{\mathbb{P}} \Bigg[\exp \Bigg(\frac{1}{\epsilon} \Big(2\log\big(\tilde{G}_i\big(\sqrt{\epsilon}W^1, \dots, \sqrt{\epsilon}W^N\big)\big) \\ - \int_0^T \sqrt{\epsilon}\dot{h}_t \mathrm{d}W_t^i + \frac{1}{2}\int_0^T (\dot{h}_t)^2 \mathrm{d}t \Big) \Bigg) \Bigg] \Bigg), \quad h \in \mathbb{H}_T \tag{10.2.12}$$

where we remark that the value of this expression does not depend on the choice of i. We then obtain the following result for $\bar{L}$ (compare with Theorem 10.2.7).

Theorem 10.2.9. *Fix $N \in \mathbb{N}$ and let Assumptions 10.2.3 and 10.2.1 hold. Assume that there exists $(u^1, \hat{u}) \in \mathbb{H}_T^2$ such that $\tilde{G}_1(u^1, \hat{u}, \dots, \hat{u}) > 0$. Then the following statements hold*

i. Let $h \in \mathbb{H}_T$ such that $\dot{h}$ is of finite variation. Then Varadhan's lemma holds for the small noise asymptotics and we can rewrite (10.2.12) *as*

$$\bar{L}(h) = \sup_{u\in\mathbb{H}_T^N} \left\{ 2\log(\tilde{G}_1(u^1, \dots, u^N)) - \int_0^T \dot{h}_t \dot{u}_t^1 \mathrm{d}t + \frac{1}{2}\int_0^T (\dot{h}_t)^2 \mathrm{d}t - \frac{1}{2}\int_0^T |\dot{u}_t|^2 \mathrm{d}t \right\}, \tag{10.2.13}$$

ii. There exists an $h^* \in \mathbb{H}_T$ which minimizes (10.2.13).

iii. Consider a simplified optimization problem

$$\sup_{u^1\in\mathbb{H}_T, \hat{u}\in\mathbb{H}_T} \left\{ 2\log(\tilde{G}_1(u^1, \hat{u}, \dots, \hat{u})) - \int_0^T (\dot{u}_t^1)^2 \mathrm{d}t - \frac{N-1}{2}\int_0^T \dot{\hat{u}}_t^2 \mathrm{d}t \right\}. \tag{10.2.14}$$

There exists a maximizer (h^{**}, u^{**}) for this problem. If

$$\bar{L}(h^{**}) = 2\log\left(\tilde{G}_1(h^{**}, u^{**}, \dots, u^{**})\right) - \int_0^T (\dot{h}_t^{**})^2 \mathrm{d}t - \frac{N-1}{2}\int_0^T (\dot{u}_t^{**})^2 \mathrm{d}t\,. \tag{10.2.15}$$

then h^{**} is asymptotically optimal and is the unique maximizer of (10.2.14)*, where we have taken $i = 1$ without loss of generality.*

The proof of this theorem is deferred to Section 10.4.1. Similarly to the previous discussion if $\log(\tilde{G}_1)$ is a concave function in $u \in \mathbb{H}_T^N$, then we know that (10.2.15) holds (this is discussed at the end of Section 10.4.1). However, in general (10.2.15) is difficult to check since, even with h^* fixed, $\bar{L}$ is still an N-dimensional optimisation problem, since (10.2.13) is supremum over $u \in \mathbb{H}_T^N$.

There is also a difficulty in quantifying how the measure change affects the propagation of chaos error i.e. a measure change that is good for the statistical error may be damaging to the propagation of chaos error. We discuss this point further in Section 10.3.

10.2.4 Computing the optimal measure change

The exponential form of the SDEs (the log-normal class) considered in [GR08] and [Rob10] allows the maximisation to be written in the form of an Euler-Lagrange equation (calculus of variations approach). Due to the more general coefficients here, we obtain a more complex interaction between the Brownian motion and the value of the SDE. Consequently we need to look towards the more general theory of optimal control to calculate the change of measure¶. Deterministic optimal control is a large subject area and one can consult [FR75] or [YZ99] for example. We recall that we are working under the $\mathbb{P}$-measure.

¶Even though we are initially dealing with SDEs, in the large deviations asymptotics, the trajectory of the Brownian motion becomes a deterministic control.

Maximum principle for Theorems 10.2.7 and 10.2.9. The maximum principle allows to translate the simplified optimization problems of Theorems 10.2.7 and 10.2.9 into boundary value problems for ODE. One can observe that we are actually interested in $\dot{u}$ rather than u, that is, in the decoupled case we can write the controlled dynamics as

$$X_t(\dot{u}) = x_0 + \int_0^t b(s, X_s(\dot{u}), \mu_s^N)\mathrm{d}s + \int_0^t \sigma \dot{u}_s \mathrm{d}s\,.$$

The theory above is for infimum while we are interested in supremum, therefore we use the fact that $\sup\{f\} = -\inf\{-f\}$.

▷ *For the decoupling algorithm* Theorem 8.6.2 yields the following equations for the adjoint function and trajectory under optimal control $\dot{u}^*$ (for a given μ^N),

$$\begin{cases} \dot{p}_t = -\partial_x b(t, X_t(\dot{u}^*), \mu_t^N)p_t\,, & p_T = \dfrac{2G'(X(\dot{u}^*))}{G(X(\dot{u}^*))}\,, \\ \dot{X}_t = b(t, X_t, \mu_t^N) + \frac{1}{2}\sigma^2 p_t\,, & X_0 = x_0\,, \end{cases} \tag{10.2.16}$$

that is, the optimal control is related to p through, $\dot{u}_t^* = \frac{1}{2}\sigma p_t$.

▷ *For the complete measure change algorithm* the argument is similar argument to the above one, but here we also need to deal with the measure term. Noting that we have two controls to optimise over (recall Theorem 10.2.9) we obtain more complex expressions. Theorem 8.6.2 yields the following system of ODEs,

$$\begin{cases} \dot{p}_t^1 = -\partial_{X^1} b(t, X_t^1, \frac{1}{N}\delta_{X_t^1} + \frac{N-1}{N}\delta_{\hat{X}_t})p_t^1 - \partial_{X^1} b(t, \hat{X}_t, \frac{1}{N}\delta_{X_t^1} + \frac{N-1}{N}\delta_{\hat{X}_t})p_t^2\,, & p_T^1 = \frac{2G'(X^1)}{G(X^1)}\,, \\ \dot{p}_t^2 = -\partial_{\hat{X}} b(t, X_t^1, \frac{1}{N}\delta_{X_t^1} + \frac{N-1}{N}\delta_{\hat{X}_t})p_t^1 - \partial_{\hat{X}} b(t, \hat{X}_t, \frac{1}{N}\delta_{X_t^1} + \frac{N-1}{N}\delta_{\hat{X}_t})p_t^2\,, & p_T^2 = 0\,, \\ \dot{X}_t^1 = b(t, X_t^1, \frac{1}{N}\delta_{X_t^1} + \frac{N-1}{N}\delta_{\hat{X}_t}) + \frac{1}{2}\sigma^2 p_t^1\,, & X_0^1 = x_0\,, \\ \dot{\hat{X}}_t = b(t, \hat{X}_t, \frac{1}{N}\delta_{X_t^1} + \frac{N-1}{N}\delta_{\hat{X}_t}) + \frac{1}{2(N-1)}\sigma^2 p_t^2\,, & \hat{X}_0 = x_0\,, \end{cases} \tag{10.2.17}$$

similarly we obtained, $\dot{u}_t^* = \frac{1}{2}\sigma p_t^1$ and $\dot{\hat{u}}_t^* = 0$ as the optimal controls. From Theorem 10.2.9 we obtain the measure change as $\dot{h} = \dot{u}$.

The difference between (10.2.16) and (10.2.17) comes from the fact that for the complete measure change we have a higher dimensional problem. That is, we have two controls and two "SDEs" thus we have more terms to optimise. Recall, when one wishes to assess asymptotic optimality, (10.2.13) is still an N-dimensional problem.

Remark 10.2.10 (Accuracy of Change of Measure)**.** *In* [GR08]*, they were able to obtain explicit solutions in certain situations, but here, due to the increase in complexity, we expect this to rarely be the case. We therefore need to set reasonable tolerances in checking whether asymptotic optimality holds.*

10.2.5 Further Discussion of Results

We have presented results for two different importance sampling algorithms. Both are interesting and have their advantages and disadvantages and we discuss this throughout the next section. However, one important aspect of our results is the fact that σ is constant. This is not ideal as many models have a non-constant σ, however, this was a necessary assumption to ensure continuity of the SDE w.r.t. the Brownian motion in high dimensions. This is required for Varadhan's lemma, hence removing such an assumption requires one to also prove Varadhan's lemma holds without the continuity condition. This is of course not a trivial problem and requires one to understand the theory of rough paths.

Another important point that is of particular interest here is whether we can use the complete measure change algorithm not to reduce the variance but to reduce the propagation of chaos error. That is, dependent on the particular measure dependence in the MV-SDE can we "push" particles to important regions and obtain an optimisation problem for that setting. We will come back to this later.

10.3 Example: Kuramoto model

The Kuramoto model is a special case of a so-called system of coupled oscillators. Such models are of particular interest in physics and are used to study many different phenomena such as active rotator systems, charge density waves and complex biological systems amongst other things, see [KŁSG02] for further details. The corresponding SDE for the Kuramoto model is

$$\mathrm{d}X_t = \left(K\int_{\mathbb{R}} \sin(y - X_t)\mu^X_{t,\mathbb{P}}(\mathrm{d}y) - \sin(X_t)\right)\mathrm{d}t + \sigma \mathrm{d}W^{\mathbb{P}}_t\,, \quad t \in [0,T], \quad X_0 = x_0\,,$$

where K is the coupling strength and σ has the physical interpretation of the temperature in the system. We consider a terminal condition $G(x) = a\exp(bx)$ (satisfying Assumption 10.2.3). Our goal is to obtain the asymptotically optimal change of measure that improves the estimation of $\mathbb{E}_{\mathbb{P}}[G(\bar{X}_T)]$.

One can see that such a model easily satisfies the assumptions required. Let us now apply the theory from the previous section to calculate the optimal change of measure. We should point out here that we do not have the concavity required for asymptotic optimality to hold automatically, therefore we need to check this condition.

By our previous discussion, to apply the decoupling algorithm here we would generate a set of N weakly interacting SDEs which we denote by $Y^{i,N}$ and approximate the original SDE by,

$$\mathrm{d}\bar{X}_t = \left(\frac{K}{N}\sum_{i=1}^{N} \sin(Y^{i,N}_t - \bar{X}_t) - \sin(\bar{X}_t)\right)\mathrm{d}t + \sigma \mathrm{d}W^{\mathbb{P}}_t\,, \quad t \in [0,T], \quad \bar{X}_0 = x_0\,.$$

Let us now apply the theory from the previous section to calculate the optimal change of measure. Our optimal control argument implies solving $\tilde{\mathbb{P}}$-a.s.

$$\text{(Decoupled)} \quad \begin{cases} \dot{p}_t = \left(\frac{K}{N}\sum_{i=1}^{N} \cos(Y^{i,N}_t - X_t) + \cos(X_t)\right)p_t\,, & p_T = 2b\,,\\ \dot{X}_t = \left(\frac{K}{N}\sum_{i=1}^{N} \sin(Y^{i,N}_t - X_t) - \sin(X_t)\right) + \frac{1}{2}\sigma^2 p_t\,, & X_0 = x_0\,. \end{cases}$$

The complete measure change algorithm yields the following system,

$$\text{(Complete)} \quad \begin{cases} \dot{p}^1_t = K\big(\frac{N-1}{N}\cos(\hat{X}_t - X^1_t) - \cos(X^1_t)\big)p^1_t - \frac{K}{N}\cos(X^1_t - \hat{X}_t)p^2_t\,, & p^1_T = 2b\,,\\ \dot{p}^2_t = -K\frac{N-1}{N}\cos(\hat{X}_t - X^1_t)p^1_t + K\big(\frac{1}{N}\cos(X^1_t - \hat{X}_t) + \cos(\hat{X}_t)\big)p^2_t\,, & p^2_T = 0\,,\\ \dot{X}^1_t = K\big(\frac{N-1}{N}\sin(\hat{X}_t - X^1_t) - \sin(X^1_t)\big) + \frac{1}{2}\sigma^2 p^1_t\,, & X^1_0 = x_0\,,\\ \dot{\hat{X}}_t = K\big(\frac{1}{N}\sin(X^1_t - \hat{X}_t) - \sin(\hat{X}_t)\big) + \frac{1}{2(N-1)}\sigma^2 p^2_t\,, & \hat{X}_0 = x_0\,, \end{cases}$$

To show the numerical advantages one can achieve by using importance sampling we consider how the time taken and the estimate given by the algorithms change with the number of particles N.

For this example we use, $T = 1$, $\bar{X}_0 = 0$, $K = 1$, $\sigma = 0.3$, $a = 0.5$ and $b = 10$. For the numerics we use an Euler scheme with step size of $\Delta t = 0.02$. The systems of equations are solved using MATLAB's `bvp4c` function. For the importance sampling, we use the particle positions from the first Monte Carlo simulation as the empirical law.

	Monte Carlo			Decoupled			Complete		
N	Payoff	Error	Time	Payoff	Error	Time	Payoff	Error	Time
1×10^3	1.5066	0.1490	3	1.5729	0.0028	9	1.5419	0.0024	3
5×10^3	1.5895	0.0626	27	1.5840	0.0013	54	1.5710	0.0013	28
1×10^4	1.6813	0.0693	76	1.5728	0.0009	153	1.5860	0.0009	75
5×10^4	1.5899	0.0200	1 025	1.5820	0.0004	2 052	1.5738	0.0004	1 062
1×10^5	1.5807	0.0176	3 433	1.5731	0.0003	6 935	1.5882	0.0003	3 644

Table 10.1: Results from standard Monte Carlo and the importance sampling algorithms. Time is measured in seconds and error refers to square root of the variance.

We recall that the decoupling importance sampling requires two runs, here we use the same N for both of these. The first note one can make is how the time scales when increasing the number of particles, namely one can truly observe the N^2 complexity[‖]. As expected the decoupling algorithm takes approximately twice as long as the standard Monte Carlo (computing the change of measure is not time consuming). Following this point we also observe that the complete measure change has roughly the same computational complexity as standard Monte Carlo. The other key point is the reduction in variance (standard error) one obtains with importance sampling. For this example we see that both importance sampling schemes reduce the variance by several orders of magnitude. Further, if one is interested in the decoupling algorithm it may be more efficient to take less simulations in the second importance sampled run. Finally, we checked the asymptotic optimality (for the decoupling) numerically and there is only a small difference between the two sides in (10.2.15), we therefore believe we are close to the optimal. Table 10.1 does show that the use of importance sampling in MV-SDEs is both viable and worthwhile.

▷ *Estimating the propagation of chaos error.* As was mentioned in the introduction, theoretically the statistical error and the propagation of chaos error converge to zero at the same rate. We now use this example to show that the statistical error dominates. Since the Euler scheme is the same in all examples we can neglect the bias caused by that. We can then decompose the error as

$$\frac{1}{N}\sum_{i=1}^{N} G(\bar{X}^{i,N}) - \mathbb{E}_{\mathbb{P}}[G(\bar{X}^{1})] = \frac{1}{N}\sum_{i=1}^{N} G(\bar{X}^{i,N}) - \mathbb{E}_{\mathbb{P}}[G(\bar{X}^{1,N})] + \mathbb{E}_{\mathbb{P}}[G(\bar{X}^{1,N})] - \mathbb{E}_{\mathbb{P}}[G(\bar{X}^{1})] \,.$$

The first difference on the RHS is the statistical error, and the second one is the propagation of chaos error. It is then clear that if one considers M realisations of $\frac{1}{N}\sum_{i=1}^{N} G(\bar{X}^{i,N})$ and takes the average this approximates $\mathbb{E}_{\mathbb{P}}[G(\bar{X}^{1,N})]$ but does not change the propagation of chaos error. Hence for large M the error reduces to the propagation of chaos error. To show the propagation of chaos error is negligible compared to the statistical error here, we repeat the simulation for $N = 5 \times 10^3$ particles, $M = 10^3$ times and we obtain an average terminal value of 1.5772 (with an average standard error of 0.06533, which agrees with the result in Table 10.1). Comparing this to the 10^5 decoupled entry (which has almost no statistical error) in Table 10.1, we can conclude the propagation of chaos error at least an order of magnitude smaller than the statistical error.

Another example: a terminal condition function with steep slope Let us consider the terminal condition $G(x) = \big(\tanh(a(x-b)) + 1\big)/2$, for a large (G can be understood as a mollified indicator function). Then $\mathbb{E}_{\mathbb{P}}[G(X_T)] \approx \mathbb{P}(X_T \geq b)$. We take the same set up as before but with $a = 15$ and $b = 1$ and note that the terminal condition for adjoint takes the form,

$$p_T = 2a\Big(1 - \tanh\big(a\big(X_T(\dot{u}^*) - b\big)\big)\Big) \,.$$

We obtain the following table (we omit the times here since they are similar).

	Monte Carlo		Decoupled		Complete	
N	Payoff (10^{-9})	Error (10^{-9})	Payoff (10^{-9})	Error (10^{-9})	Payoff (10^{-9})	Error (10^{-9})
1×10^3	1.015	0.671	3.864	0.0250	8.456	0.101
5×10^3	1.093	0.752	3.952	0.0112	5.564	0.0185
1×10^4	8.829	7.071	3.910	0.0077	32.956	0.1520
5×10^4	1.106	0.271	3.970	0.0035	2.101	0.0024
1×10^5	5.158	1.990	3.901	0.0024	16.781	0.019

Table 10.2: Results from standard Monte Carlo and the importance sampling algorithms. Note that for ease of presentation the payoff and error are all scaled to be 10^{-9} of the values presented.

The results in Table 10.2 highlight the key differences in the algorithms. Clearly this is a difficult problem for standard Monte Carlo to solve. The reason of course being that although G is mollified it still changes value quickly over a small interval. For example G at 0.25 is

[‖] Even if one is able to optimize the code somewhat, with this method we cannot escape the extra complexity arising from the particle interaction.

approximately 10^{-10}, but $G(0.5) \approx 10^{-7}$ and $G(0.75) \approx 10^{-4}$, hence a reasonably small change in the value of the SDE can influence the outcome significantly. However, for the standard Monte Carlo run, only 60 of the $100,000$ were $> 1/2$ at the terminal time and none were above $3/4$. Hence standard Monte Carlo is not giving much information about the most important region of the function.

The importance sampling schemes again give reduced errors, however, this example highlights the differences between them. Although the complete measure change does have a smaller error than standard Monte Carlo the payoff oscillates around and hence the decoupled algorithm appears to be superior since the payoffs are consistent and the error decreases in the expected manner.

▷ *Robustness of complete measure change.* The above table shows why one has to consider the effect of the measure change on the propagation of chaos error. The reason this is more prominent here than in the previous example is because the magnitude of the optimal measure change is far larger. Hence, even when we use a large number of particles they may provide a poor approximation of the law, this is where this algorithm lacks robustness.

Remark 10.3.1 (Requirement for improved simulation)**.** *It is clear from these examples that combining importance sampling with MV-SDEs can provide a major reduction in the required computational cost, namely we can achieve a smaller variance with far less simulations (and hence time). When using decoupling, unfortunately one has to approximate the law first, which is computationally expensive to do using a particle approximation. Hence, one may look towards more sophisticated simulation techniques to speed up the first run, for example* [GP15] *or towards multilevel Monte Carlo such as* [STT17]. *However, with the ability to almost eliminate the variance one should always keep in mind the benefits from importance sampling.*

10.4 Proof of Main Results

We now provide the proofs of our two main theorems. Throughout we work under the $\mathbb{P}$-measure and we omit it as a superscript in our Brownian motions. Some arguments align with those of [GR08] and we quote them where appropriate.

10.4.1 Proofs for Theorem 10.2.9

Continuity of the SDE w.r.t. Brownian motion is key as it allows to apply directly the contraction principle transferring Schilder's LDP for the Brownian motion to an LDP for the solution of the SDE; otherwise difficulties would arise when using Varadhan's lemma. Unlike the decoupled case, we will stick to Lipschitz coefficients here, the reason for this is that Lemma 10.4.3 does not generalise well for SDEs of the type (10.2.9).

Lemma 10.4.1. *Fix $N \in \mathbb{N}$, let Assumption 10.2.1 hold and let $X \in \mathbb{S}^p$ for $p \geq 2$ denote the N-dimensional strong solution to the SDE system defined in* (10.2.9). *Then X is continuous w.r.t. the set of N Brownian motions in the uniform topology.*

Proof. To show continuity in the uniform topology we consider two sets of iid Brownian motions, $W_t = (W^1_t, \dots, W^N_t)$ and $\tilde{W}_t = (\tilde{W}^1_t, \dots, \tilde{W}^N_t)$ and show continuity by analyzing the difference between, $\tilde{X}^i_t := X^i_t(\tilde{W}^1_t, \dots, \tilde{W}^N_t)$ and X^i_t with $i \in \{1, \cdots, N\}$. We have,

$$|\tilde{X}^{i,N}_t - X^{i,N}_t| \leq \int_0^t |b(s, \tilde{X}^{i,N}_s, \frac{1}{N}\sum_{j=1}^N \delta_{\tilde{X}^{j,N}_s}) - b(s, X^{i,N}_s, \frac{1}{N}\sum_{j=1}^N \delta_{X^{j,N}_s})| \mathrm{d}s + |\int_0^t \sigma \mathrm{d}\tilde{W}^i_s - \int_0^t \sigma \mathrm{d}W^i_s| \,.$$

Considering the time integral first, we can bound as follows,

$$\Big|b(s, \tilde{X}^{i,N}_s, \frac{1}{N}\sum_{j=1}^N \delta_{\tilde{X}^{j,N}_s}) - b(s, X^{i,N}_s, \frac{1}{N}\sum_{j=1}^N \delta_{X^{j,N}_s})\Big| \leq C\Big(|\tilde{X}^{i,N}_s - X^{i,N}_s| + \Big(\frac{1}{N}\sum_{j=1}^N (\tilde{X}^{j,N}_s - X^{j,N}_s)^2\Big)^{\frac{1}{2}}\Big) \,,$$

where we used the Lipschitz property and the definition of the Wasserstein-2 metric for empirical distributions (see [BJGR17], for example). Noting that for the second term,

$$\Big(\frac{1}{N}\sum_{j=1}^{N}(\tilde{X}_s^{j,N}-X_s^{j,N})^2\Big)^{\frac{1}{2}} \leq \max_{j\in\{1,\dots,N\}}|\tilde{X}_s^{j,N}-X_s^{j,N}| \leq \sum_{j=1}^{N}|\tilde{X}_s^{j,N}-X_s^{j,N}|\,.$$

Hence we can bound the drift by terms of the form $|\tilde{X}_s^{j,N}-X_s^{j,N}|$. This yields the following,

$$|\tilde{X}_t^{i,N}-X_t^{i,N}| \leq \int_0^t C\Big(|\tilde{X}_s^{i,N}-X_s^{i,N}|+\sum_{j=1}^{N}|\tilde{X}_s^{j,N}-X_s^{j,N}|\Big)\mathrm{d}s + C\sup_{0\leq s\leq t}|\tilde{W}_s^i-W_s^i|\,.$$

Taking supremums and summing over i on both sides yields,

$$\begin{aligned}\sum_{i=1}^{N}\sup_{0\leq t\leq T}|\tilde{X}_t^{i,N}-X_t^{i,N}| \leq& \int_0^T C\sum_{i=1}^{N}\sup_{0\leq t\leq s}|\tilde{X}_t^{i,N}-X_t^{i,N}|\mathrm{d}s + C\sum_{i=1}^{N}\sup_{0\leq s\leq T}|\tilde{W}_s^i-W_s^i| \\ \leq& Ce^{CT}\sum_{i=1}^{N}\sup_{0\leq s\leq T}|\tilde{W}_s^i-W_s^i|\,,\end{aligned}$$

where the final step follows from Grönwall's inequality. It is then clear that $\sum_{i=1}^{N}\sup_{0\leq s\leq T}|\tilde{W}_s^i-W_s^i|\to 0$ implies $\sum_{i=1}^{N}\sup_{0\leq t\leq T}|\tilde{X}_t^{i,N}-X_t^{i,N}|\to 0$, hence we obtain the required continuity. □

We next show that one can use Varadhan's lemma in this case.

Lemma 10.4.2. *Fix $N\in\mathbb{N}$, let $h\in\mathbb{H}_T$ and let Assumptions 10.2.3 and 10.2.1 hold.*

Then the integrability condition in Varadhan's lemma holds for (10.2.12). *Namely for* $\gamma>1$

$$\limsup_{\epsilon\to 0}\epsilon\log\Big(\mathbb{E}_{\mathbb{P}}\Big[\exp\Big(\frac{\gamma}{\epsilon}\Big(2\log\Big(\tilde{G}_1(\sqrt{\epsilon}W^1,\dots,\sqrt{\epsilon}W^N)\Big)\int_0^T\sqrt{\epsilon}\dot{h}_t\mathrm{d}W_t^1+\frac{1}{2}\int_0^T(\dot{h}_t)^2\mathrm{d}t\Big)\Big)\Big]\Big) < \infty.$$

Proof. Using that $h\in\mathbb{H}_T$ is deterministic, $\dot{h}\in L^2([0,T],\mathbb{R}^N)$ and Cauchy-Schwarz we obtain,

$$\begin{aligned}&\epsilon\log\Big(\mathbb{E}_{\mathbb{P}}\Big[\exp\Big(\frac{\gamma}{\epsilon}\Big(2\log\big(\tilde{G}_1(\sqrt{\epsilon}W^1,\dots,\sqrt{\epsilon}W^N)\big)-\int_0^T\sqrt{\epsilon}\dot{h}_t\mathrm{d}W_t^1+\frac{1}{2}\int_0^T(\dot{h}_t)^2\mathrm{d}t\Big)\Big)\Big]\Big)\\ &\leq\frac{\gamma}{2}\int_0^T(\dot{h}_t)^2\mathrm{d}t+\frac{\epsilon}{2}\log\Big(\mathbb{E}_{\mathbb{P}}\Big[\exp\Big(\frac{4\gamma}{\epsilon}\log\big(\tilde{G}_1(\sqrt{\epsilon}W^1,\dots,\sqrt{\epsilon}W^N)\big)\Big)\Big]\Big)\\ &\qquad+\frac{\epsilon}{2}\log\Big(\mathbb{E}_{\mathbb{P}}\Big[\exp\Big(-\frac{2\gamma}{\epsilon}\int_0^T\sqrt{\epsilon}\dot{h}_t\mathrm{d}W_t^1\Big)\Big]\Big)\,.\end{aligned}$$

It is then sufficient to show that the three terms are finite when we take $\limsup_{\epsilon\to 0}$. The first term is clearly finite by the conditions on h. Finiteness of the third term follows from [GR08, pg.16], namely $\forall\, i\in\{1,\dots,N\}$ the stochastic integral has the distribution $\int_0^T\dot{h}_t\mathrm{d}W_t^i\sim\mathcal{N}(0,\int_0^T(\dot{h}_t)^2\mathrm{d}t)$. Thus we obtain,

$$\limsup_{\epsilon\to 0}\frac{\epsilon}{2}\log\Big(\mathbb{E}_{\mathbb{P}}\Big[\exp\Big(-\frac{2\gamma}{\epsilon}\int_0^T\sqrt{\epsilon}\dot{h}_t\mathrm{d}W_t^1\Big)\Big]\Big)=\gamma^2\int_0^T(\dot{h}_t)^2\mathrm{d}t<\infty\,.$$

The final term to consider is the terminal term, $\log(\tilde{G}_1)$. By definition of $\tilde{G}_1$ and Assumption 10.2.3 we have,

$$\log\Big(\tilde{G}_1(\sqrt{\epsilon}W^1,\dots,\sqrt{\epsilon}W^N)\Big)\leq C_1+C_2\sup_{0\leq t\leq T}|X^{1,N}(\sqrt{\epsilon}W^1,\dots,\sqrt{\epsilon}W^N)|^{\alpha}\,.$$

Applying similar arguments as in Lemma 10.4.1 we obtain

$$
\begin{aligned}
&|X_t^{1,N}(\sqrt{\epsilon}W^1,\dots,\sqrt{\epsilon}W^N)| \\
&\leq C+\int_0^t C\Big(|X_s^{1,N}(\sqrt{\epsilon}W^1,\dots,\sqrt{\epsilon}W^N)|+\sum_{j=1}^N |X_s^{j,N}(\sqrt{\epsilon}W^1,\dots,\sqrt{\epsilon}W^N)|\Big)\mathrm{d}s+C\sqrt{\epsilon}\sup_{0\leq s\leq t}|W_s^1|\,.
\end{aligned}
$$

Noting that for $\alpha \geq 1$ and a_i nonnegative, $(\sum_{i=1}^N a_i)^\alpha \leq C^\alpha \sum_{i=1}^N a_i^\alpha$ and that the above estimate is true for any $X^{i,N}$, then taking supremums and summing over i yields,

$$
\begin{aligned}
&\sum_{i=1}^N \sup_{0\leq t\leq T}|X_t^{i,N}(\sqrt{\epsilon}W^1,\dots,\sqrt{\epsilon}W^N)|^\alpha \\
&\qquad\leq C^\alpha+\int_0^T C^\alpha\sum_{i=1}^N \sup_{0\leq t\leq s}|X_t^{i,N}(\sqrt{\epsilon}W^1,\dots,\sqrt{\epsilon}W^N)|^\alpha \mathrm{d}s+C^\alpha\sqrt{\epsilon}^\alpha\sum_{i=1}^N\sup_{0\leq s\leq T}|W_s^i|^\alpha \\
&\qquad\leq C^\alpha e^{C^\alpha}\sqrt{\epsilon}^\alpha\sum_{i=1}^N\sup_{0\leq s\leq T}|W_s^i|^\alpha\,,
\end{aligned}
$$

where the final line comes from Grönwall's inequality. It is useful for us to note this yields the bound

$$
\log\Big(\tilde{G}_1(W^1,\dots,W^N)\Big)\leq C_1+C_2\sum_{i=1}^N\sup_{0\leq s\leq T}|W_s^i|^\alpha\,. \tag{10.4.1}
$$

Using the previous results we have the following bound,

$$
\begin{aligned}
&\frac{\epsilon}{2}\log\left(\mathbb{E}_{\mathbb{P}}\left[\exp\left(\frac{4\gamma}{\epsilon}\log(\tilde{G}_1(\sqrt{\epsilon}W^1,\dots,\sqrt{\epsilon}W^N))\right)\right]\right) \\
&\qquad\leq C+\sum_{i=1}^N\frac{\epsilon C}{2}\log\left(\mathbb{E}_{\mathbb{P}}\left[\exp\left(\frac{4\gamma C^\alpha}{\epsilon^{1-\alpha/2}}\sup_{0\leq s\leq T}|W_s^i|^\alpha\right)\right]\right)\,,
\end{aligned}
$$

where we have used the independence of the Brownian motions to obtain the sum over i.

Finiteness of this term then follows by arguments similar to those in Lemma 7.6 and 7.7 in [GR08]. To conclude, we have shown that all terms are finite and the result follows. □

Before finishing the proof of Theorem 10.2.9, we note that the LDP for Brownian motion in pathspace is given by Schilder's theorem, which states that for a d-dimensional Brownian motion W, then $\sqrt{\epsilon}W$ satisfies a LDP with good rate function (see [DZ10]),

$$
I(y)=\begin{cases}\frac{1}{2}\int_0^T|\dot{y}_t|^2\mathrm{d}t & \text{, if } y\in\mathbb{H}_T^d\,,\\ \infty & \text{, if } y\in\mathbb{W}_T^d\backslash\mathbb{H}_T^d\,.\end{cases}
$$

Proof of Theorem 10.2.9. The continuity of the SDE from Lemma 10.4.1 along with existence of a unique strong solution under Assumptions 10.2.1, ensure $\tilde{G}_1$ is a continuous function under Assumption 10.2.3. By assumption, there exists a point $(u^1,\hat{u})\in\mathbb{H}_T^2$ such that $\tilde{G}(u^1,\hat{u},\dots,\hat{u})>0$, this along with (10.4.1) and recalling $\alpha<2$ we obtain the existence of maximisers by Lemma 7.1 of [GR08]. Similarly the $+\dot{h}^2$ yields existence of a minimising h for $\bar{L}$.

Moreover, continuity of $\tilde{G}$ w.r.t. the Brownian motion and finite variation of $\dot{h}$ implies the exponential term in (10.2.12) is continuous. Thus to use Varadhan's lemma we only need to check the integrability condition, which is given in Lemma 10.4.1, hence relation (10.2.13) follows.

The remaining part to be proved is that (10.2.15) implies asymptotically optimal. This essentially relies on showing that (10.2.14) is a lower bound for the RHS of (10.2.3). Using the same arguments to derive (10.2.3), one obtains the following expression for an asymptotically optimal estimator

$$\sup_{u\in\mathbb{H}_T^N}\left\{2\log(\tilde{G}_1(u^1,\dots,u^N))-\frac{1}{2}\int_0^T|\dot{u}_t|^2\mathrm{d}t\right\}.$$

It is then clear that the supremum is bounded below by the case $u^2=\dots=u^N$, which yields the expression (10.2.14).

Strict convexity along with arguments on page 18 in [GR08] yields the uniqueness which completes the proof. □

10.4.2 Proofs for Theorem 10.2.7

We recall, that due to the independence of the original particle system from the SDE in question, we work on the product of two probability spaces, consequently (since μ^N will be a "realisation" coming from the space $\tilde{\Omega}$) our results are all $\tilde{\mathbb{P}}$-a.s..

As before we need to prove that the SDE is a continuous map of the Brownian motions. We were unable to find any results for the monotone (one-sided Lipschitz), locally Lipschitz case, we therefore provide a proof of this result here (Lemma 10.4.4). The proof of this relies on the following lemma.

Lemma 10.4.3. *Let Assumption 10.2.2 hold and let $\bar{X}$ be the solution to* (10.2.4). *Then consider the following stochastic processes*

$$X_t^+:=x_0\mathbb{1}_{\{x_0\geq0\}}+\int_0^t C(|X_s^+|+1)\mathrm{d}s+\sigma\left(\sup_{0\leq s\leq t}W_s-\inf_{0\leq s\leq t}W_s\right),$$
$$X_t^-:=x_0\mathbb{1}_{\{x_0\leq0\}}-\int_0^t C(|X_s^-|+1)\mathrm{d}s+\sigma\left(\inf_{0\leq s\leq t}W_s-\sup_{0\leq s\leq t}W_s\right),$$

where C is the constant in the one-sided Lipschitz condition of b.

Then, $\forall\, t\geq0$, $X_t^-\leq\bar{X}_t\leq X_t^+$, $\mathbb{P}\otimes\tilde{\mathbb{P}}$-a.s..

Proof. Firstly, one can easily show through a standard Picard iteration argument that both $X^\pm$ have unique, progressively measurable solutions in $\mathbb{S}^2$ (see [Mao08, Section 2.3]). We argue by contradiction and show the upper bound $\bar{X}\leq X^+$, the lower bound follows by the same argument in the opposite direction. Since b is monotone (Assumption 10.2.2), we can derive the following bounds $\forall\, s\in[0,T]$ and $\mu\in\mathcal{P}_2(\mathbb{R})$,

$$b(s,x,\mu)\leq C(|x|+1)\quad\text{for } x\geq0\qquad\text{and}\qquad b(s,x,\mu)\geq -C(|x|+1)\quad\text{for } x\leq0\,.$$

Assume that there exists a time t_2 such that $\bar{X}_{t_2}>X_{t_2}^+$. If $\bar{X}_t\geq0$ for all $t\in[0,t_2]$, then,

$$\begin{aligned}X_{t_2}^+-\bar{X}_{t_2}=&x_0\mathbb{1}_{\{x_0\geq0\}}-x_0+\int_0^{t_2}C(|X_s^+|+1)-b(s,\bar{X}_s,\mu_s^N)\mathrm{d}s\\&+\sigma\big(\sup_{0\leq s\leq t_2}W_s-\inf_{0\leq s\leq t_2}W_s\big)-\sigma W_{t_2}\geq0,\end{aligned}$$

which yields a contradiction. Alternatively, let $t_1:=\max\{t\leq t_2:\bar{X}_t=0\}$. By continuity, $\bar{X}_{t_1}=0$ and so

$$\begin{aligned}X_{t_2}^+-\bar{X}_{t_2}=x_0\mathbb{1}_{\{x_0\geq0\}}+&\int_0^{t_2}C(|X_s^+|+1)\mathrm{d}s-\int_{t_1}^{t_2}b(s,\bar{X}_s,\mu_s^N)\mathrm{d}s\\&+\sigma\left(\sup_{0\leq s\leq t_2}W_s-\inf_{0\leq s\leq t_2}W_s\right)-\sigma\left(W_{t_2}-W_{t_1}\right)\geq0,\end{aligned}$$

which contradicts $\bar{X}_{t_2}>X_{t_2}^+$ and thus proves the result.

□

One can now use this lemma to prove the following result.

Lemma 10.4.4. *Let $\bar{X}$ be defined as in* (10.2.4)*, with coefficients satisfying Assumption 10.2.2, then $\bar{X}$ is a $\mathbb{P}\otimes\tilde{\mathbb{P}}$-a.s. continuous map of Brownian motion in the uniform norm.*

Proof. To prove this result we require that, if $\sup_{0\leq s\leq t}|\tilde{W}_s - W_s|\to 0$, then $\sup_{0\leq s\leq t}|\bar{X}_s(\tilde{W}) - \bar{X}_s(W)|\to 0$. We note that we work on the uniform topology and hence we may assume that all (a finite number of) Brownian motions are uniformly bounded on $[0,T]$. Lemma 10.4.3, implies that we can bound the value $\bar{X}$ takes by the processes $X^{\pm}_{\cdot}$. It is a straightforward application of Grönwall's Lemma to deduce,

$$X^+_t \leq \left(x_0\mathbb{1}_{\{x_0\geq 0\}} + Ct + \sigma\left(\sup_{0\leq s\leq t} W_s - \inf_{0\leq s\leq t} W_s\right)\right)e^{Ct}\,,$$
$$X^-_t \geq -\left(|x_0\mathbb{1}_{\{x_0\leq 0\}}| + Ct + \sigma\left|\inf_{0\leq s\leq t} W_s - \sup_{0\leq s\leq t} W_s\right|\right)e^{Ct}\,.$$

Hence we can bound the value $\bar{X}$ can take as a function of its Brownian motion (which itself is bounded by the uniform topology). Let us now consider the difference in the SDEs driven by the different Brownian motions,

$$|\bar{X}_t(\tilde{W}) - \bar{X}_t(W)| \leq \int_0^t |b(s,\bar{X}_s(\tilde{W}),\mu^N_s) - b(s,\bar{X}_s(W),\mu^N_s)|\mathrm{d}s + \sigma|\tilde{W}_t - W_t|\,.$$

By Assumption 10.2.2, b is locally Lipschitz, hence,

$$|b(s,\bar{X}_s(\tilde{W}),\mu^N_s) - b(s,\bar{X}_s(W),\mu^N_s)| \leq C(\tilde{W},W)|\bar{X}_s(\tilde{W}) - \bar{X}_s(W)|\,.$$

Noting further that $\sigma|\tilde{W}_t - W_t| \leq \sigma\sup_{0\leq s\leq t}|\tilde{W}_s - W_s|$, then by Grönwall's inequality we obtain,

$$|\bar{X}_t(\tilde{W}) - \bar{X}_t(W)| \leq \sigma\left(\sup_{0\leq s\leq t}|\tilde{W}_s - W_s|\right)e^{C(\tilde{W},W)t}\,.$$

Again, by the uniform topology, we must have $\tilde{W}$ and W bounded, thus $C(\tilde{W},W)<\infty$ and hence, $\sup_{0\leq s\leq t}|\bar{X}_s(\tilde{W}) - \bar{X}_s(W)|\to 0$ when $\sup_{0\leq s\leq t}|\tilde{W}_s - W_s|\to 0$.

□

We now prove that the uniform integrability condition still holds, namely that we can still apply Varadhan's Lemma, in both settings.

Lemma 10.4.5. *Let $h\in\mathbb{H}_T$, then under Assumption 10.2.3 and 10.2.2 the integrability condition in Varadhan's lemma holds for* (10.2.5)*. Namely, for some $\gamma>1$*

$$\limsup_{\epsilon\to 0}\epsilon\log\mathbb{E}_{\mathbb{P}\otimes\tilde{\mathbb{P}}}\Big[\exp\Big(\frac{\gamma}{\epsilon}\Big(2\log(\overline{G}(\sqrt{\epsilon}W)) - \int_0^T\sqrt{\epsilon}\dot{h}_t\mathrm{d}W_t + \frac{1}{2}\int_0^T\dot{h}^2_t\mathrm{d}t\Big)\Big)\Big|\tilde{\mathcal{F}}\Big] < \infty \quad \tilde{\mathbb{P}}\text{-a.s.}.$$

Proof. The h terms can be dealt with using the same arguments as before. The term we are interested in is the G term. Using arguments as in the proof of Lemma 10.4.2, we only need to prove the following holds,

$$\limsup_{\epsilon\to 0}\frac{\epsilon}{2}\log\left(\mathbb{E}_{\mathbb{P}\otimes\tilde{\mathbb{P}}}\left[\exp\left(\frac{4\gamma}{\epsilon}\log\left(G(\bar{X}(\sqrt{\epsilon}W))\right)\right)\Big|\tilde{\mathcal{F}}\right]\right) < \infty\,.$$

Recall that Lemma 10.4.3, yields the bound, $X^-_t \leq \bar{X}_t \leq X^+_t$, $\mathbb{P}\otimes\tilde{\mathbb{P}}$-a.s.. Hence, for $\alpha\in[1,2)$ we have the following bound $\mathbb{P}\otimes\tilde{\mathbb{P}}$-a.s.,

$$\sup_{0\leq t\leq T}|\bar{X}_t|^\alpha \leq \sup_{0\leq t\leq T}|X^+_t|^\alpha + \sup_{0\leq t\leq T}|X^-_t|^\alpha = |X^+_T|^\alpha + |X^-_T|^\alpha\,,$$

where the final equality comes from the fact $|X^{\pm}|$ are nondecreasing processes. Due to the dependence on the external measure μ^N, all of these results are $\tilde{\mathbb{P}}$-a.s., but for ease of presentation we will omit it here. Further recall that by Grönwall's lemma (or see proof of Lemma 10.4.4), we can bound the processes $|X^{\pm}|$, thus,

$$|X_T^+|^\alpha \leq C^\alpha \left(x_0^\alpha \mathbb{1}_{\{x_0 \geq 0\}} + C^\alpha + \sigma^\alpha \left(\sup_{0\leq s\leq T} W_s - \inf_{0\leq s\leq T} W_s \right)^\alpha \right) e^{C\alpha} ,$$
$$|X_T^-|^\alpha \leq C^\alpha \left(|x_0 \mathbb{1}_{\{x_0 \leq 0\}}|^\alpha + C^\alpha + \sigma^\alpha \left| \inf_{0\leq s\leq T} W_s - \sup_{0\leq s\leq T} W_s \right|^\alpha \right) e^{C\alpha} .$$

Due to the fact that $\alpha \geq 1$, and $-\inf_{0\leq s\leq T} W_s = \sup_{0\leq s\leq T} -W_s \geq 0$, we have,

$$\left| \inf_{0\leq s\leq T} W_s - \sup_{0\leq s\leq T} W_s \right|^\alpha = \left(\sup_{0\leq s\leq T} W_s - \inf_{0\leq s\leq T} W_s \right)^\alpha \leq C^\alpha \Big(\big(\sup_{0\leq s\leq T} W_s \big)^\alpha + \big(\sup_{0\leq s\leq T} -W_s \big)^\alpha \Big).$$

We express the bound w.r.t. the driving Brownian motion $\sqrt{\epsilon}W$ and obtain,

$$\sup_{0\leq t\leq T} |\bar{X}_t(\sqrt{\epsilon}W)|^\alpha \leq C^\alpha \left(|x_0|^\alpha + C^\alpha + C^\alpha \sigma^\alpha \sqrt{\epsilon}^\alpha \Big(\big(\sup_{0\leq s\leq T} W_s \big)^\alpha + \big(\sup_{0\leq s\leq T} -W_s \big)^\alpha \Big) \right) e^{C\alpha} .$$

We can simplify this further by noting,

$$\left(\sup_{0\leq s\leq T} W_s \right)^\alpha + \left(\sup_{0\leq s\leq T} -W_s \right)^\alpha \leq C^\alpha \sup_{0\leq s\leq T} |W_s|^\alpha .$$

Using these inequalities we obtain,

$$\frac{\epsilon}{2} \log \left(\mathbb{E}_{\mathbb{P}\otimes\tilde{\mathbb{P}}} \left[\exp\left(\frac{4\gamma}{\epsilon} \log(G(\bar{X}(\sqrt{\epsilon}W))) \right) \Big| \tilde{\mathcal{F}} \right] \right)$$
$$\leq \frac{\epsilon}{2} \log \left(\mathbb{E}_{\mathbb{P}\otimes\tilde{\mathbb{P}}} \left[\exp\left(\frac{4\gamma}{\epsilon} C_1 + \frac{4\gamma}{\epsilon} C_2 \left(C^\alpha \left(|x_0|^\alpha + C^\alpha + C^\alpha \sigma^\alpha \sqrt{\epsilon}^\alpha \sup_{0\leq s\leq T} |W_s|^\alpha \right) e^{C\alpha} \right) \right) \Big| \tilde{\mathcal{F}} \right] \right) .$$

By splitting up the terms in the exponential this then reduces to the problem of considering,

$$\frac{\epsilon}{2} \log \left(\mathbb{E}_{\mathbb{P}\otimes\tilde{\mathbb{P}}} \left[\exp\left(\frac{4\gamma}{\epsilon^{1-\alpha/2}} C_2 C^\alpha \sigma^\alpha \sup_{0\leq s\leq T} |W_s|^\alpha \right) \Big| \tilde{\mathcal{F}} \right] \right) .$$

One can show that this quantity is finite by following the same arguments as [GR08, pg.16]. □

We can now prove the second main theorem, the arguments follow similar lines to those we used to conclude the proof of Theorem 10.2.9.

Proof of Theorem 10.2.7. The continuity of the SDE from Lemma 10.4.4 along with existence of a unique strong solution under Assumption 10.2.2, ensure $\overline{G}$ is a $\tilde{\mathbb{P}}$-a.s. continuous function under Assumption 10.2.3. We then obtain the existence of the maximiser by Lemma 7.1 of [GR08].

Moreover, the $\tilde{\mathbb{P}}$-a.s. continuity of $\overline{G}$ w.r.t. the Brownian motion and finite variation of $\dot{h}$ implies that to use Varadhan's lemma we only need to check the integrability condition, which is given in Lemma 10.4.5. This with Lemma 7.6 in [GR08] is enough to complete the proof by arguments on page 18 in [GR08]. □

10.5 Conclusion and Outlook

For Chapter 9 we have shown how one can apply the techniques from SDEs to the MV-SDE setting and some of its pitfalls and challenges that arise. The numerical testing carried out shows that the explicit scheme yields superior results (over the implicit scheme) in general.

Although we have been able to obtain convergence for the implicit scheme it is under stronger assumptions than the explicit scheme, that being said we still observed numerical convergence in Section 9.2.3. The reason for these assumptions is that the implicit scheme is more challenging to bound than the explicit. The standard approach around this problem is to use stopping time arguments, however, as described in Remark 9.1.5 stopping times are harder to handle in the MV-SDE framework. Caution is therefore needed to account for the extra technicalities that arise.

With regard to the explicit and implicit scheme in this setting, the explicit scheme appears to be the superior choice. Namely, it has the same convergence rate as the implicit but is cheaper to compute. However, for standard SDEs the implicit scheme can handle monotone growth coefficients, that is σ need not be Lipschitz. Similar to above the existence of such MV-SDEs with monotone coefficients is not yet known but one may require an implicit type scheme to handle these equations. Therefore it is important to understand both algorithms.

We also leave open a proof for the convergence rate of the implicit scheme. Showing such a convergence rate in our framework is clearly possible but adds little in scope given the gains of the explicit over the implicit scheme. We leave the question open until a time a more resourceful implicit scheme can be designed. Moreover, to achieve such a rate potentially requires one to develop an analogous split step backward Euler scheme as in [HMS02] for MV-SDEs. Then use scheme as a stepping stone between the implicit and true and calculate the corresponding convergence rates to determine the overall convergence rate.

Another interesting area which we have not discussed is sign preservation and the impact it has on the law. For example a MV-SDE may be known to be positive, however, if the numerical scheme takes the solution into the negative region how does the law dependence influence the remaining particles? One can consider the special case of $L_b < 0$ in Assumption 8.2.1, even though the MV-SDE could have a nonnegative solution, the numerical scheme may not preserve this feature.

For Chapter 10, we have demonstrated the use of importance sampling in the MV-SDE setting and in particular how one can simulate a particle system under a measure change. Although importance sampling requires additional work, namely, calculating the measure change, we have shown that this is negligible compared to the numerical advantages one gains.

One interesting open problem this work leaves is, using importance sampling to reduce the propagation of chaos error. Namely, one can consider a hybrid decoupled-complete measure change approach whereby one uses the complete measure change algorithm to reduce the approximation of the error, then use this particle system as the “first run” in the decoupled algorithm. Due to the scaling as a function of N of the particle system, dropping the error from the law approximation by one order of magnitude implies the algorithm can be made two orders of magnitude faster for the same error.

In general in this chapter we have concentrated on increasing the practicality of MV-SDEs, whether that is providing a numerical scheme for MV-SDEs that converges under weaker assumptions or improving computational efficiency. As the field continues to be researched and more complex models are proposed, algorithms such as these will be required. In terms of finance there are some interesting measure dependent models, such as the expected value dividend model in [Bañ18]. This model has many similarities to standard Geometric Brownian Motion with dividend payments but in general the appearance of a law dependence for dividends is intuitive. For example one can set up dividend payments based on the difference between the company’s current value and its expected value i.e. pay out less when performance is below average. Of course it is common for the SDEs used to not be in the form of GBM, hence if one wishes to add a similar law dependence to these models then more general assumptions are required.

Part IV

SDEs for Solving PDEs and Branching Diffusions

Chapter 11

Preliminaries

This Part is based on the author's joint work with dos Reis [RS18].

We recall that the main goal of this part to construct a stochastic representation for first order parabolic PDEs. This problem is indeed non trivial since stochastic representations typically require second order terms. Before discussing this problem we firstly overview the literature on stochastic representations for semilinear parabolic PDEs. There are two main methods one can use here forward backward SDEs (FBSDEs) or branching diffusions. While the FBSDE methodology is well understood, the notion of branching diffusions is less well known. We therefore introduce the theory and explain the idea of branching diffusions in Chapter 12. We then look to use these ideas to construct an unbiased representation for a first order parabolic PDE in Chapter 13.

To help ensure this part is self-contained we will overview some key results. In the interest of space however, these introductions are brief and do not cover many key and interesting results within each field. We therefore encourage any interested reader to consult the references contained within for further details.

11.1 Malliavin Calculus

One of the techniques we require for our stochastic representation is Malliavin calculus, which was originally developed by Paul Malliavin in the 1970's. We shall give a short overview here, but all proofs and results can be found in texts such as [Nua06] and [DNØP09].

The theory was originally developed as an infinite dimensional[*] integration by parts techniques which could allow one to obtain smoothness results of densities for SDEs driven by Brownian motion. In fact the theory is sometimes called the stochastic calculus of variation (but this is not common). Unfortunately the scope of this theory was thought to be limited and the mathematics was too complex for the results it produced, since in most cases the density results could be obtained via Hömander's theory. However, in the mid 1980's Malliavin calculus was used to obtain an explicit formula for the process appearing in the *martingale representation theorem*, the so-called Clark-Ocone formula (see [Nua06, Proposition 1.3.14]). The martingale representation theorem is important in mathematical finance, therefore the ability to obtain an "explicit" formula is very useful. Sticking with the finance theme [FLL$^+$99] used Malliavin calculus to obtain formulas for *Greeks* (sensitivity of an option's price to various parameters). The number of areas Malliavin calculus has application in continues to grow and is too vast to mention them all here, we encourage the interested reader to consult [Nua06] and [DNØP09] for more information.

Remark 11.1.1. *It should be noted that there are different ways one can introduce the Malliavin calculus. One can approach it using Wiener-Itô chaos expansion, this is probably the most common approach and has the advantage of making some proof simpler. However, an alternative (and more intuitive approach) is to extend the notion of Fréchet derivatives to the stochastic setting*

[*] Infinite dimensional in the sense that we perform calculus on the Weiner space.

and obtain a stochastic derivative that way. An excellent overview of both approaches is given in [DNØP09, Appendix A]. *Either approach however produces the same theory.*

For our purposes we want to use Malliavin calculus as a way to rewrite derivatives w.r.t. the initial condition, this is very similar to what is done in [FLL$^+$99]. Here we present some basic (but extremely useful) results from Malliavin calculus. In order to keep our presentation as concise as possible we chose to introduce the Malliavin derivative as a stochastic derivative[†] as is done in [DNØP09, Appendix A], that is we work with the usual Wiener space and measure, hence can regard all Brownian motions as elements in $C_0([0,T])$ with values in $\mathbb{R}^d$.

Definition 11.1.2. *Assume that $F : \Omega \to \mathbb{R}$ has a directional derivative in all directions $\gamma \in \mathbb{H}^d_T$ (the Cameron Martin space) in the strong sense, that is,*

$$\mathbf{D}_\gamma F(\omega) := \lim_{\epsilon \to 0} \frac{F(\omega + \epsilon\gamma) - F(\omega)}{\epsilon}\,,$$

exists in $L^2(\Omega)$. Define $\gamma(t) = \int_0^t g(s)\mathrm{d}s$ for $g \in L^2([0,T])$ (hence $\gamma \in \mathbb{H}^d_T$) and assume in addition that there exists $\psi(t,\omega) \in L^2([0,T]\times\Omega)$ such that,

$$\mathbf{D}_\gamma F(\omega) = \int_0^T \psi(t,\omega) g(t)\mathrm{d}t \quad \textit{for all } \gamma \in \mathbb{H}^d_T\,.$$

Then we say that F is differentiable and set,

$$\mathbf{D}_t F(\omega) := \psi(t,\omega).$$

We call $\mathbf{D}_\cdot F \in L^2([0,T]\times\Omega)$ the stochastic derivative of F and the set of all differentiable random variables is denoted by $\mathcal{D}_{1,2}$.

If we now consider the space P of random variables $F : \Omega \to \mathbb{R}$ of the form,

$$F(\omega) = f\Big(\int_0^T h_1(t)\cdot \mathrm{d}W_t(\omega), \dots, \int_0^T h_n(t)\cdot \mathrm{d}W_t(\omega)\Big),$$

where f is a polynomial in each of its entries, each $h_i := (h_i^1, \dots, h_i^d) \in L^2([0,T])$ and

$$\int_0^T h_i(t)\cdot \mathrm{d}W_t(\omega) = \sum_{j=1}^d \int_0^T h_i^j(t)\cdot \mathrm{d}W_t^j(\omega)\,.$$

Note that such polynomials are referred to as Wiener polynomials and are dense in $L^2(\Omega)$. We then have the following result [DNØP09, Lemma A.12],

Lemma 11.1.3 (Chain Rule). *Let $F \in P$, then, $F \in \mathcal{D}_{1,2}$ and*

$$\mathbf{D}_t F = \sum_{i=1}^n \partial_i f\Big(\int_0^T h_1(t)\cdot \mathrm{d}W_t, \dots, \int_0^T h_n(t)\cdot \mathrm{d}W_t\Big) h_i(t)\,.$$

Reader's familiar with Malliavin calculus will realise that this is sometimes given as the definition of a derivative, see [Nua06, Definition 1.2.1]. Now we look to associate the stochastic derivative with the Malliavin derivative. To do that let us first define a norm for random variables in $\mathcal{D}_{1,2}$,

$$||F||_{1,2} = \Big(||F||^2_{L^2(\Omega)} + ||D_t F||^2_{L^2([0,T]\times\Omega)}\Big)^{1/2}\,.$$

The unfortunate result with obtaining the Malliavin derivative this way is that it is not clear that the space $\mathcal{D}_{1,2}$ is closed under this norm. We therefore define the following space.

[†] That is we look to construct the derivative using notion of Fréchet derivatives, but particularised to the setting of Wiener spaces.

Definition 11.1.4. *We define $\mathbb{D}_{1,2}$ as the closure of P w.r.t. the norm $||\cdot||_{1,2}$.*

The space $\mathbb{D}_{1,2}$ consists of random variables $F \in L^2(\Omega)$ such that there exists a sequence $F_n \in P$ which converges to F in $L^2(\Omega)$ as $n \to \infty$ and the sequence $\mathbf{D}_t F_n$ converges in $L^2([0,T] \times \Omega)$. Since the operator $\mathbf{D}_t$ is closeable on P ([DNØP09, Theorem A.14]) we can define the derivative of $F \in \mathbb{D}_{1,2}$ as,

$$D_t F := \lim_{n\to\infty} \mathbf{D}_t F_n \,.$$

This leads to the following definition.

Definition 11.1.5 (Malliavin Derivative)**.** *Let $F \in \mathbb{D}_{1,2}$, hence there exists a convergent sequence $F_n \in P$ and take,*

$$D_t F = \lim_{n\to\infty} \mathbf{D}_t F_n$$

and

$$D_\gamma F = \int_0^T D_t F \cdot g(t)\mathrm{d}t \,,$$

for all $\gamma(t) = \int_0^t g(s)\mathrm{d}s \in \mathbb{H}^d_T$. We call $D_t F$ the Malliavin derivative of F.

One can show using this definition of Malliavin derivative that the same rules for iterated Itô integrals apply, as one obtains starting from the chaos expansion approach. Hence this is indeed a consistent definition to use. The Malliavin derivative is of course only half of the story, one also requires integration. This operator is commonly referred to as the divergence operator, although it is also common to see it being referred to as the Skorohod integral. Similar to the Malliavin derivative it is possible to define the integral via Wiener-Itô chaos expansion (see [DNØP09, Chapter 2] for example). However, it can also be viewed as the adjoint operator of the Malliavin derivative via the following integration by parts [Nua06, Section 1.3].

Definition 11.1.6. *We denote by δ the adjoint of the operator D. That is, δ is an unbounded operator on $L^2([0,T] \times \Omega)$ with values in $L^2(\Omega)$ such that:*

- *The domain of δ, which we denote Domδ, is the set of $L^2([0,T] \times \Omega)$ valued random variable such that,*

  $$|\mathbb{E}[\langle DF, u\rangle]| = \Big|\mathbb{E}\Big[\int_0^T D_t F u(t)\mathrm{d}t\Big]\Big| \le C(u)||F||_{L^2(\Omega)} \,,$$

  for all $F \in \mathbb{D}_{1,2}$, C is a constant dependent on u.

- *If $u \in$ Domδ, then $\delta(u)$ is the element in $L^2(\Omega)$ characterised by,*

  $$\mathbb{E}[F\delta(u)] = \mathbb{E}[\langle DF, u\rangle] \,,$$

  for any $F \in \mathbb{D}_{1,2}$.

One can note that since D is densely defined (in $L^2([0,T] \times \Omega)$), $\delta(u)$ is unique for any $u \in$ Domδ. It is common to refer to u in Domδ as Skorohod integrable. The Skorohod integral is extremely important to obtain the properties that make Malliavin calculus so useful. The Skorohod integral is actually a generalisation of the Itô integral, however, the Skorohod integral allows for so-called *anticipating integrands*, namely integrands that are not $\mathcal{F}_t$-measurable as one requires for Itô calculus, see [NP88] or [Nua06, Section 1.3] for example. Due to the connection with the Itô integral the following notation is often used for a Skorohod integral,

$$\delta(u) =: \int_0^T u(t)\delta W_t \,.$$

Let us now state the key properties from Malliavin calculus we will use (all results are proved in [DNØP09]).

Chain Rule. Let $F \in \mathbb{D}_{1,2}$ and $g \in C_b^1(\mathbb{R})$. Then $g(F) \in \mathbb{D}_{1,2}$ and,

$$D_t g(G) = g'(G) D_t G.$$

Skorohod Extends Itô Let u be an adapted stochastic process in $L^2([0,T] \times \Omega)$. Then $u \in \mathrm{Dom}\delta$ and its Skorohod integral coincides with the Itô integral,

$$\delta(u) = \int_0^T u(t) \mathrm{d}W_t \,.$$

Integration by Parts. Let u be a Skorohod integrable stochastic process and $F \in \mathbb{D}_{1,2}$ such that the product $Fu(\cdot)$ is Skorohod integrable. Then,

$$F\delta(u) = \delta(Fu) + \int_0^T u(t) D_t F \mathrm{d}t \,.$$

First Variation and Malliavin Derivative One of the most fascinating results is the connection between the Malliavin derivative and and the so-called first variation process. Let X be an $\mathbb{R}^n$ valued SDE,

$$\mathrm{d}X_t = b(X_t)\mathrm{d}t + \sigma(X_t)\mathrm{d}W_t, \quad X_0 = x \in \mathbb{R}^n \,,$$

where $b,\ \sigma \in C_b^1$ and σ is uniformly elliptic. Then the first variation process $Y_t := \partial_x X_t$, satisfies,

$$\mathrm{d}Y_t = b'(X_t) Y_t \mathrm{d}t + \sigma'(X_t) Y_t \mathrm{d}W_t, \quad Y_0 = 1 \,.$$

This is process is discussed further in [Fri12, Section 5.5] for example. Then the first variation process is related to the Malliavin derivative as

$$D_s X_t = Y_t Y_s^{-1} \sigma(X_s) \mathbb{1}_{\{s \le t\}} \,.$$

These various properties lead to the so-called Bismut-Elworthy-Li (BEL) formula (also commonly referred to as *automatic differentiation*), see [FLL$^+$99] or [DNØP09, Theorem 4.14].

Theorem 11.1.7. *Let the assumptions on the coefficients above hold and let a be a continuous deterministic $L^2([0,T])$ function such that,*

$$\int_0^T a(t)\mathrm{d}t = 1 \,.$$

Then for any g such that $\mathbb{E}[g(X_T)^2] < \infty$ we obtain,

$$\nabla \mathbb{E}[g(X_T)] = \mathbb{E}\left[g(X_T) \int_0^T a(t) [\sigma(X_t)^{-1} Y_t]^\intercal \mathrm{d}W_t \right] \,. \tag{11.1.1}$$

There are a few interesting points to note here, firstly, the RHS does not depend on the derivative of g (and we make no assumption on g being differentiable). Furthermore, the integrand is also independent of g, hence once one can calculate the integral (common referred to as the Malliavin weight) for a given SDE, this applies to all functions g with the required integrability. Finally, the end expression does not require any Malliavin terms (D_t or δ), however, the theory is essential in order to obtain the result. This theorem will turn out to be crucial for our work on branching diffusions.

11.2 Viscosity Solutions

In Section 8.6 we considered classical (deterministic) optimal control. It turns out that the stochastic version of this theory and in particular the Dynamic Programming Principle (DPP) is extremely useful for showing that a stochastic representation solves the PDE. Recall in

Chapter 3 we started with a classical solution to the PDE then derived the corresponding stochastic representation via Itô's formula. We would now like to consider the reverse i.e. under some assumptions show that the stochastic representation is the solution to a PDE. Of course if we do not assume the PDE to have a classical solution, then one may ask does the "stochastic representation" make sense. As it turns out, one can still obtain meaning from a stochastic representation even when the PDE itself does not omit a unique classical solution, see [Pha09, Chapter 4] or [PR16, Chapter 3.8].

The solution the stochastic representation gives rise to is known as a *viscosity solution*, viscosity solutions were developed by [CL83] to analyse first order Hamilton Jacobi equations. First order Hamilton Jacobi equations were known to have weak solutions but there was a problem regarding uniqueness, as it turns out the notion of viscosity solutions allows one to obtain such uniqueness results. Although they were originally developed for first order equations this was later generalised to second order ones (see [CIL92]).

Remark 11.2.1 (Viscosity vs. weak solutions)**.** *Viscosity solutions are similar but not the same as "weak solutions" to PDEs. They are similar to weak solutions in the sense that they allow us to have a continuous function "solving" a PDE, but without the requirement on that function being differentiable. However, they were developed for optimal control type problems, where weak solutions (using integration by parts) are not convenient (showing uniqueness is an issue and nonlinearities of the solution are difficult to handle), as one uses optimal control type arguments to show stochastic representations of PDEs, viscosity solutions are the natural choice of "general" solution.*

There are various works on viscosity solutions such as [PR16, Chapter 6.5], [Pha09, Chapter 4], for a full guide on the second order case one can consult [CIL92]. These works give results on uniqueness of solutions under monotonicity assumptions however, as we are only interested in viscosity solutions as generalised solutions of PDEs we shall follow [Car08, Chapter 8.3]. Let us start by considering the following semi-linear PDE for u an $\mathbb{R}^d$ valued function defined on $[0,T]\times\mathbb{R}^l$,

$$\begin{cases}\frac{\partial u}{\partial t}(t,x)+\mathcal{L}u(t,x)+g(t,x,u(t,x),\nabla u(t,x)\sigma(t,x))=0 \quad \forall\, t\in[0,T] \text{ and } x\in\mathbb{R}^l\\ u(T,x)=\Psi(x), \quad \forall x\in\mathbb{R}^l\,,\end{cases} \tag{11.2.1}$$

where $\mathcal{L}$ is second order differential operator,

$$\mathcal{L}:=\frac{1}{2}\sum_{i,j=1}^{l}(\sigma\sigma^{\intercal}(t,x))_{i,j}\frac{\partial^2}{\partial x_i\partial x_j}+\sum_{i=1}^{l}b_i(t,x)\frac{\partial}{\partial x_i}\,.$$

In the case that b, σ, g and Ψ are not sufficiently differentiable functions, (11.2.1) will not have a classical solution. Hence we look to generalise the notion of "solution" in a sensible way (allows us to put estimates on the solution), which is goal of the viscosity solution.

We first make an assumption on the PDE, for $1\le i\le d$, the ith component of g, denoted by g_i, depends only on the ith row of $\nabla u(t,x)$. Hence we can rewrite (11.2.1) for $i=1,\dots,d$,

$$\begin{cases}\frac{\partial u_i}{\partial t}(t,x)+\mathcal{L}u_i(t,x)+g_i(t,x,u(t,x),(\nabla u(t,x))_i\sigma(t,x))=0 \quad \forall\, t\in[0,T] \text{ and } x\in\mathbb{R}^l\\ u(T,x)=\Psi(x), \quad \forall x\in\mathbb{R}^l\,,\end{cases}$$

We then have the following definition (see [Car08, Definition 8.11]).

Definition 11.2.2. *Assume $u\in C([0,T]\times\mathbb{R}^l;\mathbb{R}^d)$ and $u(T,x)=\Psi(x)$, for all $x\in\mathbb{R}^l$. For any $1\le i\le d$, and $\varphi\in C^{1,2}([0,T]\times\mathbb{R}^l)$ and $(t,x)\in[0,T]\times\mathbb{R}^l$, such that $\varphi(t,x)=u_i(t,x)$ and $u_i-\varphi$ is a local maximum at (t,x), if*

$$-\frac{\partial\varphi}{\partial t}(t,x)-\mathcal{L}\varphi(t,x)-g_i(t,x,u(t,x),\nabla\varphi(t,x)\sigma(t,x))\le 0\,,$$

then the function u is called a viscosity subsolution. Instead, let φ be such that $\varphi(t,x)=u_i(t,x)$

and $u_i - \varphi$ be a local maximum at (t, x), if

$$-\frac{\partial \varphi}{\partial t}(t,x) - \mathcal{L}\varphi(t,x) - g_i(t,x,u(t,x),\nabla\varphi(t,x)\sigma(t,x)) \geq 0\,,$$

then the function u is called a viscosity supersolution.

Moreover, u is a viscosity solution of (11.2.1) *if it is both a viscosity subsolution and viscosity supersolution.*

Further results Although we have presented viscosity solutions which are useful for semi linear parabolic PDEs. Various results exist for elliptic PDEs, see [PR16, Chapter 6.5] for example. More recent work has been carried out on extending the notion of viscosity solutions to path-dependent PDEs, see [EKT$^+$14] for further details.

11.3 (F)BSDEs

In this section we shall discuss Backward SDEs (BSDEs) and Forward BSDEs (FBSDEs). These equations are extremely important in mathematical finance since they allow one to consider incomplete markets, non linear options such as CVAs (credit valuation adjustments) and are hugely important in stochastic differential games which are fundamental in economics, see [Car08], [Car16] and [Pha09] among others. There are many texts one can follow to obtain an understanding of BSDEs such as [EKPQ97], [PR16], [Mao08] among others, however, we shall mainly follow [Car08, Chapter 8].

Let us start by defining a multidimensional BSDE [Car08, Definition 8.1]. For this we define the following spaces,

- $\mathcal{P}_n$ is the set of $\mathbb{R}^n$-valued, $\mathcal{F}_t$-progressively measurable processes on $\Omega \times [0,T]$,
- $\mathcal{S}^2_n(0,T) = \{\varphi \in \mathcal{P}_n$ with continuous paths such that $\mathbb{E}[\sup_{0\leq t\leq T}|\varphi_t|^2] < \infty\}$,
- $\mathcal{H}^2_n(0,T) = \{X \in \mathcal{P}_n$ such that $\mathbb{E}[\int_0^T |X_s|^2 \mathrm{d}s] < \infty\}$.

Definition 11.3.1. *Let $\xi_T \in L^2_T$ be an $\mathbb{R}^d$ valued terminal condition and let $g(t,\omega,y,z)$ be an $\mathbb{R}^d$ valued coefficient (commonly known as a driver), which is $\mathcal{P}_d \otimes \mathcal{B}(\mathbb{R}^d \times \mathbb{R}^{d\times l})$-measurable. A solution for the d-dimensional BSDE associated with the parameters (g, ξ_T) is a pair of progressively measurable processes $(Y_t, Z_t)_{0\leq t\leq T}$ such that the following holds,*

$$\begin{cases} Y \in S^2_d, \quad Z \in \mathcal{H}^2_{d\times l}, \\ Y_t = \xi_T + \int_t^T g(s,\omega,Y_s,Z_s)\mathrm{d}s - \int_t^T Z_s \mathrm{d}W_s, \quad 0 \leq t \leq T\,. \end{cases}$$

The differential form of this equation is,

$$-\mathrm{d}Y_t = g(t, Y_t, Z_t)\mathrm{d}t - Z_t\mathrm{d}W_t, \quad Y_T = \xi_T.$$

There are indeed some interesting differences between BSDEs and SDEs, the two main differences is that BSDE have a terminal condition ξ_T, and therefore one looks to solve at some earlier time t. The other difference is we now talk of a solution as a pair (Y, Z). The idea is, Z ensures that Y “hits” the target ξ_T at time T.

The first result on existence and uniqueness for BSDEs was given in [PP90], under the following assumption.

Assumption 11.3.2. *We assume the driver to satisfy the following:*

- $(g(t,0,0))_{t\leq T} \in \mathcal{H}^2_d$.
- *g is globally Lipschitz w.r.t. (y,z), namely there exists a constant $C \geq 0$ such that for all (y, y', z, z')*

$$|g(\omega,t,y,z) - g(\omega,t,y',z')| \leq C(|y-y'| + |z-z'|), \quad \mathrm{d}t \otimes \mathrm{d}\mathbb{P} \textit{ a.e.}$$

We then have the following result, see [Car08, Theorem 8.2].

Theorem 11.3.3. *Under Assumption 11.3.2, there exists a unique solution (Y, Z) of the BSDE with parameters (g, ξ_T).*

Similar to standard SDEs, the proof of this result can be done using either fixed point or Picard approximation. One can find explicit solutions in some cases and also prove existence of solutions with more general assumptions. However, we do not discuss this further here and instead move onto an important class of BSDEs.

As it turns out, if the BSDE coefficients are deterministic functions of a diffusion process, then the solution (Y, Z) is also a deterministic function of that process. This also has connections to semi-linear PDEs but this will be discussed later. The framework here is that the terminal condition of the BSDE (sometimes referred to as a Markovian BSDE) is governed by some underlying diffusions process $(X_s^{t,x})_{0\le s\le T}$, which is the strong solution to the SDE,

$$\begin{cases} \mathrm{d}X_s^{t,x} = b(s, X_s^{t,x})\mathrm{d}s + \sigma(s, X_s^{t,x})\mathrm{d}W_s, & t \le s \le T \\ X_s^{t,x} = x & 0 \le s < t\,. \end{cases} \tag{11.3.1}$$

Further, for any[‡] $(t, x) \in [0, T] \times \mathbb{R}^l$, we denote by $(Y_s^{t,x}, Z_s^{t,x})_{0\le s\le T}$, the solution to the BSDE,

$$\begin{cases} -\mathrm{d}Y_s^{t,x} = g(s, X_s^{t,x}, Y_s^{t,x}, Z_s^{t,x})\mathbb{1}_{\{s\ge t\}}\mathrm{d}s - Z_s^{t,x}\mathrm{d}W_s \\ Y_T^{t,x} = \Psi(X_T^{t,x})\,. \end{cases} \tag{11.3.2}$$

The system (11.3.1) and (11.3.2) is referred to as a *forward backward stochastic differential equation* (FBSDE). FBSDEs are extremely important in finance, such as option pricing where the forward process denotes the stock and the terminal value is the payoff of the option. In order for the system to have reasonable estimates we require some assumption on the forward process.

Assumption 11.3.4. *We assume the following on the coefficients of* (11.3.1).

- *b and σ are uniformly Lipschitz continuous w.r.t. x.*
- *there exists a constant $c > 0$ s.t. for any (s, x),*

$$|b(s, x)| + |\sigma(s, x)| \le c(1 + |x|)\,.$$

One also requires assumptions on the coefficients of the BSDE[§], we then make the equivalent assumption to Assumption 11.3.2.

Assumption 11.3.5. *Let the following hold.*

- *The function $g : [0, T] \times \mathbb{R}^l \times \mathbb{R}^d \times \mathbb{R}^{d\times l} \to \mathbb{R}^d$ is uniformly Lipschitz in (y, z) with Lipschitz constant C i.e.,*

$$|g(s, x, y_1, z_1) - g(s, x, y_2, z_2)| \le C(|y_1 - y_2| + |z_1 - z_2|)\,.$$

- *There exists a constant C such that for a real constant $p \ge 1/2$,*

$$|g(s, x, 0, 0)| + |\Psi(x)| \le C(1 + |x|^p)\,.$$

We then have the following existence result for an FBSDE.

Theorem 11.3.6. *Let Assumptions 11.3.4 and 11.3.5 hold, there exists two measurable deterministic functions $u(t, x)$ and $d(t, x)$ such that the solution $(Y_s^{t,x}, Z_s^{t,x})$ of BSDE* (11.3.2) *is given by,*

$$\forall s \le T,\;\; Y_s^{t,x} = u(s, X_s^{t,x}), \quad \textit{and} \quad Z_s^{t,x} = d(s, X_s^{t,x})\sigma(s, X_s^{t,x}), \quad \mathrm{d}s \otimes \mathrm{d}\mathbb{P}\text{-}a.e.$$

[‡] Note that we are taking the diffusion process to be l dimensional, hence, σ is a square matrix.

[§] One should note that while we previously allowed g to be random, it is now a completely deterministic function.

Furthermore, for any $\mathcal{F}_t$-measurable random variable $\chi \in L^2$, the solution $(Y_s^{t,\chi}, Z_s^{t,\chi})_{s\geq t}$ is given by $(u(s, X_s^{t,\chi}), d(s, X_s^{t,\chi}))_{s\geq t}$.

Remark 11.3.7 (More general assumptions)**.** *Although Assumption 11.3.5 is standard, these can be generalised, see* [Car08, Chapter 8] *and references therein. Such results are very important in the context of finance, however, in order to keep this section concise we do not discuss them.*

As we shall discuss, the FBSDE system offers us a way to solve a semi-linear PDE. Indeed the forward diffusion process is the driving SDE in Feynman-Kac representation, and the solution to the PDE is the Y process of the corresponding BSDE.

11.3.1 Solving a Non Linear Second Order PDE: FBSDEs

Now that we have introduced the FBSDE framework let us give the main result on FBSDEs solving semi linear PDEs. Although it is possible to start with a classical solution to the PDE and show that the FBSDE solves this PDE just by applying Itô, typically one only has knowledge of the coefficients of the PDE (or FBSDE) and these do not necessarily lead to a classical solution. Therefore, one must generalise the notion of solution and it turns out that viscosity solutions (see Section 11.2) are ideal in this setting. Recall in Definition 11.2.2 we require the viscosity solution to be continuous, hence we require the following assumption.

Assumption 11.3.8. *The mapping $x \to (g(t,x,0,0), \Psi(x))$ is continuous for all $t \in [0,T]$.*

We may now state the crucial result, see [Car08, Theorem 8.12].

Theorem 11.3.9. *Under Assumptions 11.3.4, 11.3.5 and 11.3.8, the function $u(t,x) := Y_t^{t,x}$ is a viscosity solution to the PDE* (11.2.1) *and grows at most polynomially at infinity.*

This result states that under the appropriate Lipschitz assumptions the FBSDE can tell us about the solution of the PDE. This result was originally given in [PP92], however, by assuming more regularity [PP92] proved another interesting connection, [Car08, Theorem 8.14].

Theorem 11.3.10. *Let Assumptions 11.3.4, 11.3.5 and 11.3.8 hold and additionally assume that b, σ, g and Ψ are three times continuously differentiable with bounded derivatives w.r.t. x. Then,*

1. *If u belongs to $C^{1,2}([0,T] \times \mathbb{R}^l, \mathbb{R}^d)$ is a classical solution to* (11.2.1)*, then the couple $(u(s, X_s^{t,x}), \nabla u(s, X_s^{t,x})\sigma(s, X_s^{t,x}))$ is a solution to* (11.3.2) *in the time interval $[t,T]$. In addition, for any $t \leq T$, $u(t,x) = Y_t^{t,x}$.*
2. *If $(Y_s^{t,x}, Z_s^{t,x})$ is the unique solution of the* (11.3.2)*, then $u(t,x) := Y_t^{t,x}$, $0 \leq t \leq T$, $x \in \mathbb{R}^l$ belongs to $C^{1,2}([0,T] \times \mathbb{R}^l, \mathbb{R}^d)$ and it is a classical solution of* (11.2.1).

The second result in this theorem is interesting since it uses the FBSDE system in order to obtain results for the solution of the PDE. However, at a practical level, Theorem 11.3.9 is the more important result.

Numerics for FBSDEs

Recall from Chapter 3 one of the main motivations for obtaining a stochastic representation is that it allows one to make numerical gains over deterministic algorithms through probabilistic domain decomposition. At this point we have only described results at the theoretical level, namely, there exists a solution to the FBSDE system (and hence the PDE). The goal here is to discuss methods to (approximate) obtain such a solution.

Similar to standard SDEs, in general the BSDE (or FBSDE) solution is unknown and therefore one must look towards numerical schemes in order to estimate it. The simulation of the forward diffusion is well known (see [KP11] for example), as it turns out however, simulation of the BSDE part is much more involved. Even though the work on the connection with PDEs was carried out in the early 90's it was more than 10 years later before a viable numerical scheme was proposed. There were some algorithms developed beforehand, but these required high regularity and or required estimation of several integrals. The first real breakthrough was given in [Zha04] and this was soon followed up by [BT04], which is now the standard numerical scheme for BSDE simulation.

To see why BSDEs are more challenging, let us consider the naive Euler discretisation of (11.3.2). That is consider some partition $0 = t_0 < \cdots < t_M = T$, for some integer $M > 0$, denoting the discretised version of the BSDE as $\bar{Y}$ and $\bar{Z}$, we have terminal value $\bar{Y}_{t_M} = \Psi(\bar{X}_T)$, and recursively define (the backwards step) by,

$$\bar{Y}_{t_i} - \bar{Y}_{t_{i-1}} = -g(t_{i-1}, \bar{X}_{t_{i-1}}, \bar{Y}_{t_{i-1}}, \bar{Z}_{t_{i-1}})(t_i - t_{i-1}) + \bar{Z}_{t_{i-1}}(W_{t_i} - W_{t_{i-1}})\,.$$

There are several problems with this algorithm, firstly it is an implicit scheme (recall we wish to solve for $\bar{Y}_{t_{i-1}}$) but this is more not ideal, than catastrophic. Secondly, we only have one equation and two unknowns i.e. $\bar{Y}$ and $\bar{Z}$. Finally (and possibly most importantly) $\bar{Y}$ and $\bar{Z}$ are not adapted to the filtration in this scheme since time t_{i-1} depends on the Brownian motion at time t_i. The solution to this is not obvious, essentially by either multiplying by the Brownian increment or not and taking conditional expectations we can split the backward scheme into two steps. The corresponding numerical scheme for the BSDE is, set $\bar{Y}_{t_M} = \Psi(\bar{X}_T)$, then for $i = 1, \dots, M-1$ we recursively define (see [BT04]),

$$\begin{aligned}
\bar{Z}_{t_{i-1}} &= (t_i - t_{i-1})^{-1}\mathbb{E}[\bar{Y}_{t_i}(W_{t_i} - W_{t_{i-1}})|\mathcal{F}_{t_{i-1}}]\,,\\
\bar{Y}_{t_{i-1}} &= \mathbb{E}[\bar{Y}_{t_i}|\mathcal{F}_{t_{i-1}}] + g(t_{i-1}, \bar{X}_{t_{i-1}}, \bar{Y}_{t_{i-1}}, \bar{Z}_{t_{i-1}})(t_i - t_{i-1})\,.
\end{aligned}$$

This scheme solves the issues above, namely we now have two equations for our two unknowns, but crucially due to the conditional expectations we have that the scheme is adapted. Therefore we have a viable numerical scheme for the BSDE.

The only issue one now faces is how to approximate the conditional expectations. It turns out there are many ways in which to do this. The original approach is so-called optimal basis fitting as suggested in [LS01], a different approach is to use Malliavin calculus as in [BET04]. More sophisticated approaches for BSDE simulation are overviewed in [CM10].

The simulation of BSDEs is still a very active area of research, more recent methods and relaxing of assumptions can be found in [BTW17], [HJK16], [GT16] and [LdRS15] among several others. There has also been work into an extension of BSDEs to so-called two BSDEs (or second order BSDEs) as considered in [CSTV07] and more recently [PT$^+$15]. These have connections to fully nonlinear PDEs, however, we not discuss these further here but encourage interested readers to consult these texts for more details.

Although BSDEs have many applications and are quite general, the requirement to go backwards makes algorithms computationally expensive. Indeed, the fact a conditional expectation appears in the discretisation can be viewed as a Monte Carlo within a Monte Carlo. Of course the reason for the conditional expectation in the scheme is in order to ensure adaptedness of the solution. Ideally we could capture semi-linear PDEs (as we can with BSDEs) but without the expensive backward stepping (as is the case with a diffusion process), we present such a method in the next chapter.

Chapter 12

Branching Diffusion Overview

In Chapter 3 we saw how one can derive stochastic representations for PDEs (the Feynman-Kac formula) and gave motivation as to why this is an important result. One issue we faced however, was that the SDE methodology broke down when we tried to extend the theory to so-called semi-linear PDEs. In Section 11.3 we described the BSDE methodology and showed that although the approach is general, the requirement to go backwards makes numerical schemes expensive. The goal of this chapter is to present (in some sense) the best of both worlds, namely, an algorithm capable of handling nonlinear PDEs but is also completely "forward". This is the so-called *branching diffusion* approach. It should be noted that nothing appearing in this chapter is new, the goal is mainly to present the theory of branching diffusions and this will help describe some of the key features of the method presented in Chapter 13.

12.1 Solving a Non Linear Second Order PDE: Branching Diffusions

As mentioned in Chapter 3, the theory of branching diffusions can be traced as far back as the 1960's with [Sko64] and [Wat65], and to solving a specific PDE (the KPP equation) in [McK75]. One issue with the method in these papers however, was the requirement to have specific forms of equation, therefore it was not a general tool. However, the far more recent works of [RRM10] and [HL12] provided a simple idea as to how such results could be extended to PDEs with polynomial dependence in the solution[*].

Following that [HLTT14] provided a rigorous framework and used branching diffusions as a way to solve BSDEs with polynomial drivers w.r.t. the solution Y (and hence were viscosity solutions of semi-linear PDEs with polynomial dependence in the solution). Building on this work [HLOT^{+}16] used Malliavin calculus to show branching diffusions give rise to a viscosity solution with polynomial dependence in the solution and the gradient of the solution. Therefore creating a completely forward stochastic representation of a semi-linear PDE.

One of the issues surrounding branching diffusions is that proofs showing the stochastic process is integrable and a viscosity solution for the PDE are very technical. Most of these technical details will be covered in Chapter 13, therefore the main goal of this section is to present the idea of branching diffusions and "why" they solve nonlinear PDEs. As a consequence, until we come to state any theorems we will assume the PDE has a unique classical solution with bounded derivatives.

[*] The fact that we are limited to polynomial dependence is not as restrictive as it first sounds since many functions can be well approximated by polynomials.

12.1.1 Semi-Classical Case

We start by considering the following PDE

$$\begin{cases}\partial_t u + \mathcal{L}u + c\left(\sum_{i=0}^{M}\alpha_i u^i - u\right) = 0 & (t,x)\in[0,T)\times\mathbb{R}\,,\\ u(T,\cdot)=\Psi(\cdot) & x\in\mathbb{R}\,,\end{cases} \tag{12.1.1}$$

where c is a positive constant and for all $i\in\{0,\dots,M\}$, $\alpha_i\geq 0$, $\sum_{i=0}^{M}\alpha_i=1$ and

$$\mathcal{L} := \frac{1}{2}\sigma(x)^2\frac{\partial^2}{\partial x^2} + b(x)\frac{\partial}{\partial x}\,.$$

Of course, one could just use BSDEs but the goal here is to obtain a completely forward representation of this solution. Under the same arguments as presented in Chapter 3, assuming u exists and satisfies (12.1.1), we obtain,

$$\mathrm{d}u(t,X_t^{s,y}) = -\,c\left(\sum_{i=0}^{M}\alpha_i u(t,X_t^{s,y})^i - u(t,X_t^{s,y})\right)\mathrm{d}t + \sigma(X_t^{s,y})\partial_x u(t,X_t^{s,y})\mathrm{d}W_t\,,$$

with driving SDE,

$$\mathrm{d}X_t^{s,y} = b(X_t^{s,y})\mathrm{d}t + \sigma(X_t^{s,y})\mathrm{d}W_t,\quad X_s^{s,y}=y\,.$$

We now consider the transform $\tilde{u}(t,x)=u(t,x)e^{-ct}$, hence we obtain the following SDE for $\tilde{u}$,

$$\begin{aligned}\mathrm{d}\tilde{u}(t,X_t^{s,y}) &= -c\exp(-ct)u(t,X_t^{s,y})\mathrm{d}t + \exp(-ct)\mathrm{d}u(t,X_t^{s,y})\\ &= -c\exp(-ct)u(X_t^{s,y})\mathrm{d}t + -c\exp(-ct)\Big(\sum_{i=0}^{M}\alpha_i u(t,X_t^{s,y})^i - u(t,X_t^{s,y})\Big)\mathrm{d}t\\ &\quad + \exp(-ct)\sigma(X_t^{s,y})\partial_x u(t,X_t^{s,y})\mathrm{d}W_t\\ &= \exp(-ct)\left[-c\sum_{i=0}^{M}\alpha_i u(t,X_t^{s,y})^i\mathrm{d}t + \sigma(X_t^{y})\partial_x u(X_t^{s,y})\mathrm{d}W_t\right].\end{aligned}$$

Hence, solving for $\tilde{u}$ between s and T we obtain,

$$\tilde{u}(T,X_T^{s,y}) = \tilde{u}(s,X_s^{s,y}) - \int_s^T e^{-ct}c\sum_{i=0}^{M}\alpha_i u(t,X_t^{s,y})^i\mathrm{d}t + \int_s^T e^{-ct}\sigma(X_t^{s,y})\partial_x u(X_t^{s,y})\mathrm{d}W_t\,.$$

Converting back into u and rearranging for $u(s,y)$ yields,

$$u(s,X_s^{s,y}) = u(T,X_T^{s,y})e^{-c(T-s)} + e^{cs}\int_s^T e^{-ct}c\sum_{i=0}^{M}\alpha_i u(t,X_t^{s,y})^i\mathrm{d}t - e^{cs}\int_s^T e^{-ct}\sigma(X_t^{s,y})\partial_x u(X_t^{s,y})\mathrm{d}W_t\,.$$

Taking conditional expectation (again w.r.t. the filtration at time s),

$$u(s,y) = \mathbb{E}_{s,y}\left[u(s,X_s^{s,y})\right] = \mathbb{E}_{s,y}\left[\Psi(X_T^{s,y})e^{-c(T-s)} + e^{cs}\int_s^T e^{-ct}c\sum_{i=0}^{M}\alpha_i u(t,X_t^{s,y})^i\mathrm{d}t\right].$$

At first glance it appears we have not improved the situation from Chapter 3, i.e. we still have a dependency on u inside the expectation, however, we will be able to remove it after making some observations. To make notation easier let us take $s=0$. Namely, we have a terminal value PDE and we want to know the solution at $(0,y)$. The stochastic representation of this is then,

$$u(0,y) = \mathbb{E}_{0,y}\left[\Psi(X_T^{0,y})e^{-cT} + \int_0^T e^{-ct}c\sum_{i=0}^{M}\alpha_i u(t,X_t^{0,y})^i\mathrm{d}t\right]. \tag{12.1.2}$$

We are now in the position to introduce the notion of "particles", at first this seems like an abstract construction but it will become more intuitive as we manipulate the expectation. Let us denote by τ the lifetime of the particle starting at time 0, we take τ as a r.v. such that $\tau \sim \text{Exp}(c)$ (exponential random variable with intensity c). We then make the following observation,

$$e^{-cT} = \mathbb{P}[\tau \geq T] = \mathbb{E}[\mathbb{1}_{\{\tau \geq T\}}] , \qquad \text{the CDF of } \tau \sim \text{Exp}(c).$$

Substituting this into the first term in the expectation we can rewrite,

$$\mathbb{E}_{0,y}\left[\Psi(X_T^{0,y})\mathbb{E}[\mathbb{1}_{\{\tau\geq T\}}]\right] .$$

One must be careful here with the expectations, moreover, we change our σ-algebra here to $\mathcal{F} = \mathcal{F}^X \otimes \mathcal{F}^\tau = \sigma(X_s : \ s \leq T, \tau)$, where $\mathcal{F}^X$ is the original σ-algebra from the process X and $\mathcal{F}^\tau$ is the σ-algebra from the exponential random variable. Taking $\tau \perp\!\!\!\perp W$ (independent), we use the independence to rewrite the conditional expectation as,

$$\begin{aligned}\mathbb{E}_{0,y}\left[\Psi(X_T^{0,y})\mathbb{E}[\mathbb{1}_{\{\tau\geq T\}}]\right] =& \mathbb{E}_{0,y}\left[\Psi(X_T^{0,y})\mathbb{E}[\mathbb{1}_{\{\tau\geq T\}}|\mathcal{F}^X]\right] \quad \text{(independence)}\\ =& \mathbb{E}_{0,y}\left[\mathbb{E}[\Psi(X_T^{0,y})\mathbb{1}_{\{\tau\geq T\}}|\mathcal{F}^X]\right] \quad \text{(taking out what is known, in reverse)}\\ =& \mathbb{E}_{0,y}\left[\Psi(X_T^{0,y})\mathbb{1}_{\{\tau\geq T\}}\right] \quad \text{(tower property)} .\end{aligned}$$

We now want to do the same for the second term in (12.1.2). This is more complicated since X appears inside an integral, but a similar argument can be used. Firstly, we would like to use the tower property to condition with $\mathcal{F}^X$, therefore we can treat X as deterministic inside the integral. Then noting, ce^{-ct} is the density function of an exponential r.v. with intensity c, we can introduce the indicator $\mathbb{1}_{\{t<T\}}$ and we obtain an expectation over $\tau \sim \text{Exp}(c)$ (same as previous). Thus we can write,

$$\begin{aligned}\mathbb{E}_{0,y}\left[\int_0^T e^{-ct}c\sum_{i=0}^M \alpha_i u(t, X_t^{0,y})^i \mathrm{d}t\right] &= \mathbb{E}_{0,y}\left[\mathbb{E}\left[\int_0^\infty e^{-ct}c\mathbb{1}_{\{t<T\}}\sum_{i=0}^M \alpha_i u(t, X_t^{0,y})^i \mathrm{d}t \middle| \mathcal{F}^X\right]\right]\\ &= \mathbb{E}_{0,y}\left[\mathbb{E}\left[\mathbb{1}_{\{\tau<T\}}\sum_{i=0}^M \alpha_i u(\tau, X_\tau^{0,y})^i \middle| \mathcal{F}^X\right]\right]\\ &= \mathbb{E}_{0,y}\left[\mathbb{1}_{\{\tau<T\}}\sum_{i=0}^M \alpha_i u(\tau, X_\tau^{0,y})^i\right] .\end{aligned}$$

Therefore, we can rewrite the solution to u as

$$u(0,y) = \mathbb{E}_{0,y}\left[\Psi(X_T^{0,y})\mathbb{1}_{\{\tau\geq T\}} + \sum_{i=0}^M \alpha_i u(\tau, X_\tau^{0,y})^i \mathbb{1}_{\{\tau<T\}}\right] .$$

This gives an intuitive statement, either the particle lives beyond the terminal time $\tau \geq T$ or it "branches" $\tau < T$. The case where the particle lives to time T is trivial, the key point to discuss now is what happens in the case of a branching (the event $\tau < T$).

To do this one considers a new r.v. $I \perp\!\!\!\perp \tau, W$ such that $I \sim \text{MN}(\alpha)$, (I is multinomial parameter $\alpha = (\alpha_1, \dots, \alpha_M) \in \mathbb{R}^M$, that is $\mathbb{P}(I = i) = \alpha_i$ for $i \in \{1, \dots, M\}$). Recall that I is well defined in the sense that α defines a probability distribution. We then expand our σ-algebra again, hence define $\mathcal{F} = \mathcal{F}^X \otimes \mathcal{F}^\tau \otimes \mathcal{F}^I = \sigma(X_s : \ s \leq T, \tau, I)$ (σ-algebra generated by the random variables). For ease of writing we will denote by $\mathbb{E}_{\tau, X_\tau}[\cdot] = \mathbb{E}[\cdot|\sigma(\tau, X_s : \ s \leq \tau)]$, note that the σ-algebra generated by X_τ works here since we are considering τ as well. Again, by a Tower property argument and noting the indicator $\mathbb{1}_{\{\tau<T\}}$ is measurable w.r.t. $\sigma(\tau, X_\tau)$ we can write,

$$\begin{aligned}\mathbb{E}_{0,y}\left[\sum_{i=0}^{M}\alpha_i u(\tau, X_\tau^{0,y})^i \mathbb{1}_{\{\tau<T\}}\right] &= \mathbb{E}_{0,y}\left[\mathbb{1}_{\{\tau<T\}}\mathbb{E}_{\tau,X_\tau}\left[\sum_{i=0}^{M}\alpha_i u(\tau, X_\tau^{0,y})^i\right]\right] \\ &= \mathbb{E}_{0,y}\left[u(\tau, X_\tau^{0,y})^I \mathbb{1}_{\{\tau<T\}}\right] .\end{aligned}$$

For the next part it will be beneficial for us to keep the conditional expectation there, thus we obtain,

$$u(0,y) = \mathbb{E}_{0,y}\left[\Psi(X_T^{0,y})\mathbb{1}_{\{\tau\geq T\}} + \mathbb{1}_{\{\tau<T\}}\mathbb{E}_{\tau,X_\tau}\left[u(\tau, X_\tau^{0,y})^I\right]\right] .$$

Due to the flow property of SDEs we can write, $\mathbb{E}_{\tau,X_\tau}\left[u(\tau, X_\tau^{0,y})^I\right] = \mathbb{E}_{\tau,X_\tau}\left[u(\tau, X_\tau^{\tau,X_\tau})^I\right]$. Conditional on τ and X_τ, the function $u(\tau, X_\tau^{\tau,X_\tau})$ has a known representation, hence we can substitute this in,

$$\begin{aligned}&\mathbb{E}_{0,y}\Bigg[\Psi(X_T^{0,y})\mathbb{1}_{\{\tau\geq T\}} \\ &+ \mathbb{1}_{\{\tau<T\}}\mathbb{E}_{\tau,X_\tau}\left[\left(\Psi(X_T^{\tau,X_\tau})\mathbb{1}_{\{\tau+\tau'\geq T\}} + \mathbb{1}_{\{\tau+\tau'<T\}}\mathbb{E}_{\tau+\tau',X_{\tau+\tau'}}\left[u(\tau+\tau', X_{\tau+\tau'}^{\tau+\tau',X_{\tau+\tau'}})^{I'}\right]\right)^I\right]\Bigg],\end{aligned}$$

for τ' and I' identically distributed to τ and I but independent. This finally allows us to obtain a representation in terms of the particles.Previously we had a term u^I, hence we consider this as I particles. By taking τ' independent for each particle $1, \dots, I$, and each Brownian motion also independent, then every particle is independent of each other[†]. Therefore we are able to take the product outside the expectation, hence, $\mathbb{E}_{\tau,X_\tau}\left[u(\tau, X_\tau^{\tau,X_\tau})^I\right] = \mathbb{E}_{\tau,X_\tau}\left[u(\tau, X_\tau^{\tau,X_\tau})\right]^I$.

Before writing the full expression we need to relabel to have a consistent notation. Define τ_k as the survival time of the kth particle, where k is the index of that particle and $k-$ the index of the parent particle[‡], the index $k=1$ is reserved for the initial particle. Further, denote by T_k the total time the particle survived to (thus $T_1 = \tau_1$), hence we have the recursive relation $T_k = (T_{k-} + \tau_k) \wedge T$. Finally, denote by $\mathcal{K}_t$ the indices of the alive particles at time t, note since I can equal zero it is possible for a particle to have no descendants, i.e. dies. We obtain the following relation,

$$\begin{aligned}u(0,y) = \mathbb{E}_{0,y}\Bigg[&\Psi(X_T^{0,y})\mathbb{1}_{\{\tau_1\geq T\}} + \\ &\mathbb{1}_{\{\tau_1<T\}}\prod_{k\in\mathcal{K}_{\tau_1}}\mathbb{E}_{\tau_1,X_{\tau_1}}\left[\Psi(X_T^{\tau_1,X_{\tau_1},k})\mathbb{1}_{\{T_k\geq T\}} + \mathbb{1}_{\{T_k<T\}}\mathbb{E}_{T_k,X_{T_k}}\left[u(T_k, X_{T_k}^{\tau_1,X_{\tau_1},k})\right]^I\right]\Bigg],\end{aligned}$$

where we see $k \in \mathcal{K}_{\tau_1}$ corresponds to I, we also introduce k into X to denote the particle index. This relation gives us the notion of the branching process, namely, if $\tau_1 < T$, we have $I = |\mathcal{K}_{\tau_1}|$ particles (independent of each other), all starting at time τ_1 and point X_{τ_1}. To help keep the notation as simple as possible we will adopt the notation $X_t^{0,y,k}$ to denote the time t location of particle k (with $k \in \mathcal{K}_t$) where the initial particle started at point y at time 0. Therefore, $X_t^{0,y,k} = X_t^{T_{k-},X_{T_{k-}},k}$.

Remark 12.1.1 (A Comment on the Filtration)**.** *The goal here is to give a detailed introduction that allows us to show how the branching representation arises. The problem we now face is, as soon as we introduce multiple (independent) particles, each of these require their own Brownian motion, exponential stopping time and (if need be) their own multinomial branching. Consequently it is common practice to state at the start that we take the filtration defined with as many random variables as we need and therefore, never need to constantly redefine our filtration.*

Clearly each $\tau_k > 0$ ($\mathbb{P}$-a.s.), hence by continuing to expand (substitute for) u^I we will eventually have the set $\mathcal{K}_T$ (i.e. continue to expand until every particle hits the terminal

[†] Independence here is w.r.t. the conditional expectation, since all I particles start at time τ and position X_τ.

[‡] Parent particles need not be unique i.e. $k-$ may be the same for many different k

boundary T) creating a completely independent set of random variables for each particle. For $k \in \mathcal{K}_T$ (thus $\mathbb{1}_{\{T_k<T\}} = 0$), we obtain the following stochastic representation,

$$u(0,y) = \mathbb{E}_{0,y}\Bigg[\prod_{k\in\mathcal{K}_T} \Psi(X_T^{0,y,k})\Bigg].$$

One notes that if the case $\tau_1 \geq T$ occurs, then $\mathcal{K}_T = \{1\}$. The key result here however, is that we have a representation for the nonlinear PDE but it is dependent only on the terminal value and the SDE (diffusion process). Therefore the representation is a completely forward representation.

To give an intuitive explanation for the algorithm we consider the case where the non linear term in the one dimensional PDE is $\frac{1}{2}(1+u(t,x)^2)$. Therefore if an event occurs before terminal time T, our particle can either "die", which is the realisation $I = 0$ or branch into two independent particles $I = 2$. We give a potential realisation in Figure 12.1.

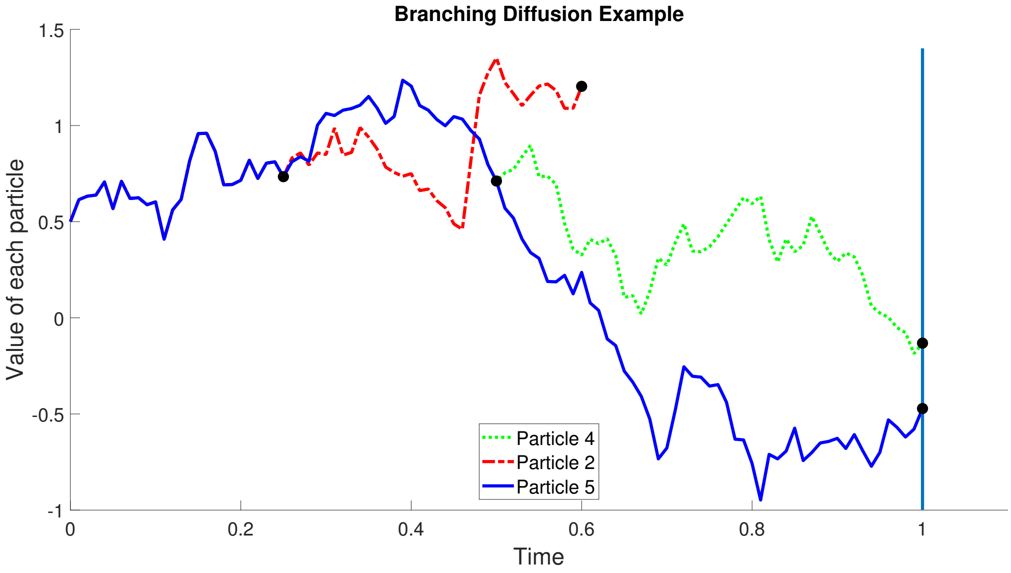


Figure 12.1: The figure shows a potential realisation of a branching diffusion. We start with an initial particle and at $\tau_1 = 0.25$ branches into two particles. One of those particles branches into two again at time 0.5, while the other one dies at time 0.6 (hence $\tau_2 = 0.35$ and $\tau_3 = 0.25$). The remaining two particles then go on to hit the terminal boundary (so τ_4 and τ_5 are both > 0.5). We also have $\mathcal{K}_T = \{4,5\}$.

Remark 12.1.2 (Computational Complexity)**.** *This expression yields the desired* forward *representation. However, there are two important quantities that one must check are finite. Since we will use Monte Carlo to solve this we require finite variance, but also due to particles branching we need to ensure the number of particles remains finite. These will be some of the main considerations through the remainder of this chapter (and Chapter 13).*

12.1.2 Marked Branching Diffusion Case

The semi classical case provides an idea of how (forward) stochastic representations can be built for non-linear PDEs. The main issue with the semi-classical case though is, the condition $\sum_{i=0}^{M} \alpha_i = 1$ with $\alpha_i \geq 0$ is too restrictive for most applications. As it turns out though, it is not too difficult to construct an algorithm that generalises this, as considered in [HL12] and [RRM10]. Following that a solid mathematical framework for showing the connection between branching diffusions, BSDEs and PDEs was given in [HLTT14]. The intuition here is almost identical to the previous section, hence we only point out the key step.

Let us generalise to the following PDE

$$\begin{cases} \partial_t u + \mathcal{L}u + c\left(\sum_{i=0}^M \alpha_i u^i - u\right) = 0 & (t,x) \in [0,T) \times \mathbb{R}\,, \\ u(T,\cdot) = \Psi(\cdot) & x \in \mathbb{R}\,, \end{cases}$$

where again c is a positive constant. However, we do not assume α is a constant and α_i sum to one. As we are only giving intuition at this point let us again assume that the PDE has a classical bounded solution so the arguments follow. The idea now is to define a probability distribution p, hence $\sum_{i=0}^M p_i = 1$ and for all i, $p_i \geq 0$, with the restriction, if there exists $(t,x) \in [0,T) \times \mathbb{R}$, such that $\alpha_i(t,x) \neq 0$, then $p_i > 0$. It is then a simple task to rewrite the PDE as,

$$\begin{cases} \partial_t u + \mathcal{L}u + c\big(\sum_{i=0}^M p_i \dfrac{\alpha_i}{p_i} u^i - u\big) = 0 & (t,x) \in [0,T) \times \mathbb{R}\,, \\ u(T,\cdot) = \Psi(\cdot) & x \in \mathbb{R}\,, \end{cases} \tag{12.1.3}$$

Following the exact same argument as in the semi-classical case we can arrive at the following stochastic representation for the solution at $u(0,y)$,

$$u(0,y) = \mathbb{E}_{0,y}\left[\Psi(X_T^{0,y}) \mathbb{1}_{\{\tau \geq T\}} + \sum_{i=0}^M p_i \frac{\alpha_i(\tau, X_\tau^{0,y})}{p_i} u(\tau, X_\tau^{0,y})^i \mathbb{1}_{\{\tau < T\}} \right].$$

As p is a probability distribution, then we can again introduce a multinomial distribution on this sum. Namely, we say the process branches in I particles where $I \sim \text{MN}(p)$, but now such a branching carries the weight $\frac{\alpha_I}{p_I}$. This make the representation more complex since each "branching" carries its own weight, however, one can capture many more PDEs with this generalisation.

Remark 12.1.3 (Choice of p)**.** *In principle the choice of p doesn't make a difference (provided it satisfies the conditions) however, like importance sampling, there are "good" choices of p in terms of variance reduction. This is discussed further in* [HL12].

The assumptions required for the convergence are given in [HLTT14], however, this algorithm has been generalised further. Therefore to avoid unnecessary overlap we leave out stating the main result and instead focus on the more general case.

12.1.3 Age Marked Branching Diffusions

The final increase in generality we make are so-called *age-dependent marked branching diffusions*, as developed in [HLOT$^+$16]. We now consider the more general semi-linear PDE for functions[§] $u : [0,T] \times \mathbb{R}^d \to \mathbb{R}$,

$$\begin{cases} \partial_t u + \mathcal{L}u + f(t,x,u,Du) = 0 & (t,x) \in [0,T) \times \mathbb{R}^d\,, \\ u(T,\cdot) = \Psi(\cdot) & x \in \mathbb{R}^d\,, \end{cases} \tag{12.1.4}$$

where Du is the gradient of u and $\mathcal{L}$ is the standard diffusion generator,

$$\mathcal{L} := \frac{1}{2}\sum_{i,j=1}^d \left(\sigma\sigma^\intercal(t,x)\right)_{i,j} \frac{\partial^2}{\partial x_i \partial x_j} + \sum_{i=1}^d b_i(x) \frac{\partial}{\partial x_i}\,.$$

Let us consider nonnegative integer m, then $L \subset \mathbb{N}^{m+1}$ and a sequence of functions $(c_\ell)_{\ell \in L}$ and $(a_i)_{i=1,\dots,m}$, where $c_\ell : [0,T] \times \mathbb{R}^d \to \mathbb{R}$ and $a_i : [0,T] \times \mathbb{R}^d \to \mathbb{R}$. For every $\ell = (\ell_0, \ell_1, \dots, \ell_m) \in L$, denote $|\ell| := \sum_{i=0}^m \ell_i$. Then the function f (the *driver* in the context of BSDEs), is a function

[§] Observe that taking u a function in $\mathbb{R}$ implies the notion of viscosity solutions follows easily, see Definition 11.2.2.

$f : [0,T] \times \mathbb{R}^d \times \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}$, of the form,

$$f(t,x,y,z) = \sum_{\ell=(\ell_0,\ell_1,\dots,\ell_m)\in L} c_\ell(t,x) y^{\ell_0} \prod_{i=1}^{m} (a_i(t,x)\cdot z)^{\ell_i} . \tag{12.1.5}$$

Hence using this representation of f we can consider semilinear PDEs with polynomial dependence in u and Du. The exact dependence is determined by the set L, for example, consider the case $m = 1$, and $L = \{(0,0),(1,1)\}$, then,

$$\sum_{\ell=(\ell_0,\ell_1,\dots,\ell_m)\in L} c_\ell(t,x) y^{\ell_0} \prod_{i=1}^{m} (a_i(t,x)\cdot z)^{\ell_i} = c_{(0,0)}(t,x) + c_{(1,1)}(t,x) y (a_1(t,x)\cdot z) .$$

The second term in the context of (12.1.4) is a term of the form "uDu", we have seen in previous sections how one can use branching arguments to handle nonlinearities in u, what we shall now explain is how to deal with terms of the form Du. Similar to previous we start by presenting the intuition behind the idea under "nice" assumptions. Following that though we shall state the representation result precisely.

For ease of presentation for the moment we drop the general case of f as in (12.1.5) and follow the example as presented in [HLOT$^+$16], which is the second order extension to Burgers equation (in one dimension)

$$\begin{cases} \partial_t u + \frac{1}{2}\partial_{xx}u + \frac{1}{2}\big(u(t,x)^2 + u\partial_x u(t,x)\big) = 0 & (t,x) \in [0,T) \times \mathbb{R} , \\ u(T,\cdot) = \Psi(\cdot) & x \in \mathbb{R} , \end{cases} \tag{12.1.6}$$

linking this with previous we have $L = \{(2,0),(1,1)\}$. Assuming u, b and σ are all sufficiently differentiable and bounded, then the solution to this PDE can be represented via,

$$u(0,y) = \mathbb{E}\Big[\Psi(W_T + y) + \int_0^T \frac{1}{2}\big(u(s, y + W_s)^2 + u\partial_x u(s, W_s + y)\big)\mathrm{d}s\Big] ,$$

where we have used that we can solve the SDE exactly in this case, i.e. $X_s^{0,y} = y + W_s$. By introducing some probability distribution ρ on $\mathbb{R}_+$ (previously we used exponential but this turns out not to be the best choice), then denote by $\overline{F}(T) := \int_T^\infty \rho(s)\mathrm{d}s$, the so-called *survival function*. Following the same arguments as previous we can introduce τ_1 from distribution described by ρ and I a multinomial over the two events (since the cardinality of L is two and $I \in L$), hence the representation is,

$$u(0,y) = \mathbb{E}\Big[\frac{\Psi(W_T + y)}{\overline{F}(T)} \mathbb{1}_{\{\tau_1 \geq T\}} + \mathbb{1}_{\{\tau_1 < T\}} \frac{1}{\rho(\tau_1)} \big(u^{I_1(1)} \partial_x u^{I_1(2)}(\tau_1, W_{\tau_1} + y)\big)\Big] .$$

In the case $I_1 = (2,0)$ we deal with this term in the same fashion as detailed above. The more interesting case is $I_1 = (1,1)$, by independence of particles arguments we can write this term as,

$$\begin{aligned} u\partial_x u(\tau_1, W_{\tau_1} + y) = \mathbb{E}\Big[&\frac{\Psi(W_T^{1,2} + y)}{\overline{F}(T - \tau_1)} \mathbb{1}_{\{T_2 \geq T\}} + \mathbb{1}_{\{T_2 < T\}} \frac{1}{\rho(\tau_2)} u^{I_2(1)} \partial_x u^{I_2(2)}(T_2, W_{T_2}^{1,2} + y)\Big|\mathcal{F}_1\Big] \\ &\times \partial_x \mathbb{E}\Big[\frac{\Psi(W_T^{1,3} + y)}{\overline{F}(T - \tau_1)} \mathbb{1}_{\{T_3 \geq T\}} + \mathbb{1}_{\{T_3 < T\}} \frac{1}{\rho(\tau_3)} u^{I_3(1)} \partial_x u^{I_3(2)}(T_3, W_{T_3}^{1,3} + y)\Big|\mathcal{F}_1\Big], \end{aligned}$$

where $\mathcal{F}_1$ is the filtration generated up to time τ_1 (we shall make this more precise later). Note that we have adopted the same notation for X and put k into the superscript of W to denote each particles independent Brownian motion. To deal with the ∂_x term we can use the so-called *automatic differentiation* technique as shown in (11.1.1). In this case it is simple since our first variation process is just a constant equal to one (and $\sigma = 1$), thus we obtain,

$$\begin{aligned}
&\partial_x \mathbb{E}\Big[\frac{\Psi(W_T^{1,3}+y)}{\overline{F}(T-\tau_1)}\mathbb{1}_{\{T_3\geq T\}} + \mathbb{1}_{\{T_3<T\}}\frac{1}{\rho(\tau_3)}u^{I_3(1)}\partial_x u^{I_3(2)}(T_3, W_{T_3}^{1,3}+y)\Big|\mathcal{F}_1\Big] \\
&= \mathbb{E}\Big[\Big(\frac{\Psi(W_T^{1,3}+y)}{\overline{F}(T-\tau_1)}\mathbb{1}_{\{T_3\geq T\}} + \mathbb{1}_{\{T_3<T\}}\frac{1}{\rho(\tau_3)}u^{I_3(1)}\partial_x u^{I_3(2)}(T_3, W_{T_3}^{1,3}+y)\Big)\frac{W_{T_3}^{1,3}-W_{\tau_1}^{1}}{\tau_3}\Big|\mathcal{F}_1\Big].
\end{aligned}$$

Hence we obtain the following representation,

$$\begin{aligned}
u(0,y) =& \mathbb{E}\Bigg[\frac{\Psi(W_T+y)}{\overline{F}(T)}\mathbb{1}_{\{\tau_1\geq T\}} + \mathbb{1}_{\{\tau_1<T\}}\frac{1}{\rho(\tau_1)}\mathbb{E}\Bigg[\\
&\Big(\frac{\Psi(W_T^{1,2}+y)}{\overline{F}(T-\tau_1)}\mathbb{1}_{\{T_2\geq T\}} + \mathbb{1}_{\{T_2<T\}}\frac{\partial_x u^{I_2(2)}(T_2, W_{T_2}^{1,2}+y)}{\rho(\tau_2)}u^{I_2(1)}\Big) \\
&\times\Big(\frac{\Psi(W_T^{1,3}+y)}{\overline{F}(T-\tau_1)}\mathbb{1}_{\{T_3\geq T\}} + \mathbb{1}_{\{T_3<T\}}\frac{u^{I_3(1)}\partial_x u^{I_3(2)}(T_3, W_{T_3}^{1,3}+y)}{\rho(\tau_3)}\Big)\Big(\frac{W_{T_3}^{1,3}-W_{\tau_1}^{1}}{\tau_3}\Big)^{I_1(2)}\Big|\mathcal{F}_1\Bigg]\Bigg].
\end{aligned}$$

Hence we return back to the original particle representation that is the above has the same form as u^2, except now any derivative term carries the *Malliavin weight*. Again we can then continue to expand out the u terms to remove the dependence on it.

General Branching Representation

Now that we have given the intuition on how the branching representation arises we state the general case as considered in [HLOT$^+$16], indeed all proofs and results in this section are given in [HLOT$^+$16]. The volume of notation surrounding branching diffusions can make the results difficult to understand. In order to make the result as readable as possible we stick to a similar notation to [HLOT$^+$16] and for completeness we restate all our notation. For the moment we take the initial condition of the system as $(0, x_0)$, but we will look to generalise this later.

In order to construct the branching diffusion process we first of all consider a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ which also contains,

- a sequence of i.i.d. positive random variables $(\tau_{m,n})_{m,n\geq 1}$ of density ρ,
- a sequence of i.i.d. random integers $(I_{m,n})_{m,n\geq 1}$, with distribution $\mathbb{P}(I_m = \ell) = p_\ell$, $\ell \in L$.

Additionally we take the two sequences $(\tau_{m,n})_{m,n\geq 1}$ and $(I_{m,n})_{m,n\geq 1}$ as independent. Using these sequences we can construct the branching process as (see [HLOT$^+$16, Section 2.2])

1. We firstly start with a particle (which has mark 0) indexed by (1) and whose lifetime (arrival time) is denoted by $T_1 := \tau^{1,1} \wedge T$.

2. We take $k = (k_1, \ldots, k_n) \in \mathbb{N}^n$ as a particle of generation n, namely there has been $n-1$ "branching events" to construct this particle. The arrival time of this particle is denoted by T_k. Let us assume that $T_k < T$ (hence a branching event occurs), with event of type $I^{n,\pi_n(k)}$, where π_n is an injection from $\mathbb{N}^n$ to $\mathbb{N}$. Therefore at time T_k, the particle branches into $|I^{n,\pi_n(k)}|$ particles (which are of generation $n+1$), and each particle is indexed by $(k_1, \ldots, k_n, i)$ for $i = 1, \ldots, |I^{n,\pi_n(k)}|$. Note that the case $|I^{n,\pi_n(k)}| = 0$ is the case where the particle dies with no offspring (although this does not imply it does not contribute to the expectation). To make our notation simpler let us define $I_k := I^{n,\pi_n(k)}$.

3. Each $I_k = (\ell_0, \ell_1, \ldots, \ell_m)$, which implies we have $|\ell|$ offspring particles from particle k. Of these offspring particles we mark the ℓ_0 by 0, the ℓ_1 by 1 etc, until every offspring has a mark¶ associated to it.

¶Note the distinction between "index" and "mark". The index identifies the particle, hence every particle has its own unique index. While the mark will be used to identify the Malliavin weight, therefore multiple particles can have the same mark.

4. For a particle of index $k = (k_1, \dots, k_n, k_{n+1})$ (an $n+1$ generation particle) we denote by $k-$ its "parent" i.e. $k- = (k_1, \dots, k_n)$, which is clearly not unique in general. Hence the arrival time of particle k can be written w.r.t. its parent as, $T_k := (T_{k-} + \tau^{n+1, \pi_{n+1}(k)}) \wedge T$. Alternatively, T_{k-} is the "birth time" of particle k and the arrival time of particle $k-$. For the initial particle, $k = (1)$, we adopt the convention $k- = \emptyset$ and $T_\emptyset = 0$.

Although the situation is more complex here, the picture of what is going on is still the same as Figure 12.1. The only difference is we must record additional information at each branching event so one can calculate the value of the mark (weight). In order to be able to state the representation we introduce the notation, θ_k denotes the mark associated to the particle with index k, and the set of particles of generation n alive at time t by,

$$\mathcal{K}_t^n := \begin{cases} \{k \ : \ \text{of generation } n \text{ s.t. } T_{k-} \leq t < T_k\}, & \text{when } t \in [0, T) \\ \{k \ : \ \text{of generation } n \text{ s.t. } T_k = T\}, & \text{when } t = T\,. \end{cases}$$

Associated to this we denote by $\overline{\mathcal{K}}_t^n := \cup_{s \leq t} \mathcal{K}_s^n$ as the set of all particles of generation n alive before time t, $\mathcal{K}_t := \cup_{n \geq 1} \mathcal{K}_t^n$ as the set of particles (of any generation) alive in the system at time t and finally $\overline{\mathcal{K}}_t := \cup_{n \geq 1} \overline{\mathcal{K}}_t^n$ as the set of all particles that have lived before time t.

Clearly, in this process a particle can produce several other particles we need to check the system has finite computational cost (namely the number of particles remains finite), this is given by [HLOT$^+$16, Proposition 2.3].

Proposition 12.1.4. *Assume that $\sum_{\ell \in L} |\ell| p_\ell < \infty$. Then the branching process is well defined on $[0, T]$, namely $\overline{\mathcal{K}}_T$ is a.s. finite.*

Of course as we have discussed, τ and I only constitute some features of the branching diffusion, we finally look to equip each particle with a Brownian motion[‖]. Similar to our other random variables we consider a sequence of independent d-dimensional Brownian motions $(W^{m,n})_{m,n \geq 1}$, which are also independent of $(\tau_{m,n}, I_{m,n})_{m,n \geq 1}$. Define for the initial particle $W_t^{(1)} = \Delta W_t^{(1)} := W_t^{1,1}$, for $t \in [0, T_{(1)}]$, then for subsequent generation particles $k = (k_1, \dots, k_n) \in \overline{\mathcal{K}}_T \backslash \{(1)\}$ we define the Brownian motion as, $W_t^k := W_{T_{k-}}^{k-} + \Delta W_{t - T_{k-}}^k$, where $\Delta W_{t - T_{k-}}^k := W_{t - T_{k-}}^{n, \pi_n(k)}$, for $t \in [T_{k-}, T_k]$. Recall $W_\cdot^{n, \pi_n(k)}$ is independent of all other Brownian motions.

With these notions, $(W_\cdot^k)_{k \in \overline{\mathcal{K}}_T}$ is referred to as a *branching Brownian motion* and for each $k \in \overline{\mathcal{K}}_T$, we associate with this Brownian motion a diffusion process $(X_t^k)_{t \in [T_{k-}, T_k]}$ via,

$$X_t^k = X_{T_{k-}}^{k-} + \int_{T_{k-}}^t b(s, X_s^k) \mathrm{d}s + \int_{T_{k-}}^t \sigma(s, X_s^k) \mathrm{d}W_s^k, \quad t \in [T_{k-}, T_k], \quad \mathbb{P}\text{-a.s.},$$

where for particle (1), the initial condition is $X_0^{(1)} = x_0$, where $x_0 \in \mathbb{R}^d$ is some constant. This process $(X_\cdot^k)_{k \in \overline{\mathcal{K}}_T}$ is our marked branching diffusion process.

For the branching diffusion process we make the following assumptions ([HLOT$^+$16, Assumptions 3.1 and 3.2]).

Assumption 12.1.5.

1. *The probability mass function $(p_\ell)_{\ell \in L}$ satisfies $p_\ell > 0$ for every $\ell \in L$, and $\sum_{\ell \in L} |\ell| p_\ell < \infty$. Moreover, the density function $\rho : \mathbb{R}_+ \to \mathbb{R}_+$ is continuous, with $\rho(t) > 0$ for all $t \in [0, T]$ and $\overline{F}(T) = \int_T^\infty \rho(s) \mathrm{d}s > 0$.*

2. *The coefficient in SDE $(b, \sigma) : [0, T] \times \mathbb{R}^d \to \mathbb{R}^d \times \mathbb{R}^{d \times d}$ are bounded continuous and Lipschitz in x.*

3. *$c_\ell : [0, T] \times \mathbb{R}^d \to \mathbb{R}$ and $a_i : [0, T] \times \mathbb{R}^d \to \mathbb{R}^d$ are bounded continuous.*

[‖] Without the Brownian motion, what we have described is an age-dependent marked branching process. Adding the Brownian motion turns the branching process into a branching diffusion.

The next assumption we make is for the *automatic differentiation* on the diffusion $\overline{X}_s^{t,x}$ defined by,

$$\overline{X}_s^{t,x} = x + \int_t^s b(r, \overline{X}_r^{t,x})\mathrm{d}r + \int_t^s \sigma(r, \overline{X}_r^{t,x})\mathrm{d}W_r, \quad s \in [t,T],$$

where W is d-dimensional Brownian motion.

Assumption 12.1.6. *There exists a measurable functional $\overline{\mathcal{W}}(t, s, x, (W_r - W_s)_{r\in[t,s]})$ which is a continuous mapping $(t,x) \to \overline{\mathcal{W}}(t, s, x, (W_r - W_t)_{r\in[t,s]})$, and for any $s \in [t,T]$ and bounded measurable function $\phi : \mathbb{R}^d \to \mathbb{R}$,*

$$\partial_x \mathbb{E}[\phi(\overline{X}_s^{t,x})] = \mathbb{E}[\phi(\overline{X}_s^{t,x})\overline{\mathcal{W}}(t, s, x, (W_r - W_s)_{r\in[t,s]})].$$

As discussed in Section 11.1, such a $\overline{\mathcal{W}}$ can be obtained for an appropriate diffusion process via Malliavin calculus, although it should be noted that such $\overline{\mathcal{W}}$ are not unique.

We are now in a position to state the main representation as in [HLOT$^+$16]. Let us denote by

$$\mathcal{W}_k := \mathbb{1}_{\{\theta_k=0\}} + \mathbb{1}_{\{\theta_k\neq 0\}} a_{\theta_k}(T_{k-}, X^k_{T_{k-}}) \cdot \overline{\mathcal{W}}(T_{k-}, T_k, X^k_{T_{k-}}, \Delta W^k_\cdot).$$

Then for a smooth function $u \in C^{1,2}([0,T] \times \mathbb{R}^d)$ we denote by ψ_n its nth generation representation, that is,

$$\begin{aligned}\psi_n :=& \Big[\prod_{k\in\cup_{j=1}^n \mathcal{K}_T^j} \frac{\Psi(X_T^k) - \Psi(X^k_{T_{k-}})\mathbb{1}_{\{\theta_k\neq 0\}}}{\overline{F}(\Delta T_k)}\mathcal{W}_k\Big]\Big[\prod_{k\in\cup_{j=1}^n \{\overline{\mathcal{K}}_T^j \setminus \mathcal{K}_T^j\}} \frac{c_{I_k}(T_k, X^k_{T_k})}{p_{I_k}}\frac{\mathcal{W}_k}{\rho(\Delta T_k)}\Big] \\ & \Big[\prod_{k\in\overline{\mathcal{K}}_T^{n+1}} \big(\mathbb{1}_{\{\theta_k=0\}}u + \sum_{i=1}^m \mathbb{1}_{\{\theta_k=i\}} a_i \cdot Du\big)(T_{k-}, X^k_{T_{k-}})\Big],\end{aligned} \tag{12.1.7}$$

for all $n \geq 1$ and the limit $n \to \infty$ as,

$$\psi := \Big[\prod_{k\in\mathcal{K}_T^j} \frac{\Psi(X_T^k) - \Psi(X^k_{T_{k-}})\mathbb{1}_{\{\theta_k\neq 0\}}}{\overline{F}(\Delta T_k)}\mathcal{W}_k\Big]\Big[\prod_{k\in\{\overline{\mathcal{K}}_T^j \setminus \mathcal{K}_T^j\}} \frac{c_{I_k}(T_k, X^k_{T_k})}{p_{I_k}}\frac{\mathcal{W}_k}{\rho(\Delta T_k)}\Big]. \tag{12.1.8}$$

As described in [HLOT$^+$16], ψ and ψ_n are defined for the initial condition $(0, x_0)$, however, the same argument applies if we perturb to the initial condition to (t,x), to make the initial condition explicit we put it in as a superscript, e.g. $\psi^{t,x}$. The final process we introduce is similar to ψ but has a different control variate,

$$\tilde{\psi} := \Big[\prod_{k\in\mathcal{K}_T^j} \frac{\Psi(X_T^k) - \Psi(X^k_{T_{k-}})\mathbb{1}_{\{\theta_k\neq 0 \text{ or } k=(1)\}}}{\overline{F}(\Delta T_k)}\mathcal{W}_k\Big]\Big[\prod_{k\in\{\overline{\mathcal{K}}_T^j \setminus \mathcal{K}_T^j\}} \frac{c_{I_k}(T_k, X^k_{T_k})}{p_{I_k}}\frac{\mathcal{W}_k}{\rho(\Delta T_k)}\Big].$$

We can now use these processes to state the main branching diffusion result, [HLOT$^+$16, Theorem 3.5].

Theorem 12.1.7. *Let Assumptions 12.1.5 and 12.1.6 hold, moreover assume for all $(t,x) \in [0,T] \times \mathbb{R}^d$ there exists $\epsilon > 0$ such that,*

$$(\psi^{s,y})_{(s,y)\in B_\epsilon(t,x)} \quad \textit{and} \quad \big(\tilde{\psi}^{s,y}\overline{\mathcal{W}}(s, T^s_{(1)}, y, \Delta W^{s,(1)}_\cdot)\big)_{(s,y)\in B_\epsilon(t,x)}$$

are uniformly integrable, where $B_\epsilon(t,x) := \{(s,y) \in [0,T] \times \mathbb{R}^d : |s-t| + |x-y| \leq \epsilon\}$.

Then the function $u(t,x) := \mathbb{E}[\psi^{t,x}]$ is a continuous viscosity solution of the semilinear PDE (12.1.4). *Moreover, u has a continuous first derivative in space.*

Although the above theorem is general, the additional assumptions are not easy to verify, moreover, we do not say anything about ψ being square integrable (a requirement to use Monte

Carlo). To address these issues [HLOT$^+$16] consider the following more explicit assumptions ([HLOT$^+$16, Assumption 3.6 and 3.10])

Assumption 12.1.8. *The coefficients b and σ are bounded continuous, with bounded continuous gradients, Db and $D\sigma$. Moreover, σ is uniformly elliptic.*

Let $n \geq 1$ and $q > 1$, and denote by L_Ψ the Lipschitz constant of Ψ. We then denote by $B_0^\infty(L_\Psi) := \{(x_1, \ldots, x_d) \in \mathbb{R}^d : |x_i| \leq L_\Psi$ for $i = 1, \ldots, d\}$ and $\overline{\mathcal{W}}_{t,x,s} := \overline{\mathcal{W}}(t, s, x, (W_r - W_t)_{r\in[t,s]})$. We then introduce the following constants $C_{1,q}$ and $C_{2,q}$,

$$C_{1,q} := |\Psi|_\infty^q \vee \sup_{0\leq t<s\leq T, x\in\mathbb{R}^d, i=1,\ldots,m, a_0\in B_0^\infty(L_\Psi)} \mathbb{E}\Big[\Big|\big(a_0 \cdot (\overline{X}_s^{t,x} - x)\big)\big(a_i(t,x)\cdot \overline{\mathcal{W}}_{t,x,s}\big)\Big|^q\Big],$$

and

$$C_{2,q} := |\Psi|_\infty^q \vee \sup_{0\leq t<s\leq T, x\in\mathbb{R}^d, i=1,\ldots,m} \mathbb{E}\Big[\Big|\big(\sqrt{s-t}\, a_i(t,x)\cdot \overline{\mathcal{W}}_{t,x,s}\big)\Big|^q\Big].$$

Then the following modifications to these constants,

$$\hat{C}_{1,q} := \frac{C_{1,q}}{\overline{F}(T)^{q-1}}, \quad \text{and} \quad \hat{C}_{2,q} := C_{2,q} \sup_{\ell\in L, t\in(0,T]} \Big(\frac{|c_\ell|_\infty}{p_\ell}\frac{t^{-q/(2(q-1))}}{\rho(t)}\Big)^{q-1}.$$

With the above constants we make the assumption.

Assumption 12.1.9. *There exists a $q > 1$, such that one of the following two holds*

1. *Both $C_{1,q}(1/\overline{F}(T))^q$ and $\sup_{\ell\in L, t\in(0,T]} C_{2,q}\Big(\frac{|c_\ell|_\infty}{p_\ell}\frac{1}{\sqrt{t}\rho(t)}\Big)^q$ are bounded by 1.*
2. $$T < \int_{\hat{C}_{1,q}} \big(\hat{C}_{2,q}\sum_{\ell\in L}|c_\ell|_\infty x^{|\ell|}\big)^{-1}\mathrm{d}x.$$

This then leads to the following result (see [HLOT$^+$16, Theorem 3.12])

Theorem 12.1.10. *Let Assumptions 12.1.5, 12.1.8 and 12.1.9 hold. Then the following is true.*

1. *Assumption 12.1.6 holds and $\big(\psi^{t,x}, \tilde{\psi}^{t,x}\overline{\mathcal{W}}(t, T_{(1)}^t, x, \Delta W_{(1)}^t)\big)_{(t,x)\in[0,T]\times\mathbb{R}^d}$ is uniformly integrable. Hence the representation $u(t,x) := \mathbb{E}[\psi^{t,x}]$ is a viscosity solution of* (12.1.4).
2. *If one further has that Assumption 12.1.9 holds for some $q \geq 2$, then $\mathbb{E}[|\psi^{t,x}|^2] < \infty$.*

We therefore have conditions under which the branching representation is square integrable. Although Assumption 12.1.9 seems somewhat arbitrary it can be viewed as a "small maturity" or "small nonlinearity" restriction.

Branching diffusions is still a very active area of research as papers look to extend the idea to more settings. For examples on this one can consult, [AC17], [CT17], [HLT18], [War18] among others for such results.

Remark 12.1.11. *The one drawback for branching diffusions over BSDEs is that one requires stricter assumptions. The overall message is that BSDEs will work in more settings, but when branching diffusions work, they are much more computationally efficient.*

12.2 Introduction to Unbiased Simulation of SDEs

Recently the ideas from branching diffusions have been used to create a method to approximate an expected value of an SDE with no error coming from the time discretisation, see [HLTT17] and [DOW17]. The discretisation error is in some sense a "worse" error than the statistical error, the reason is that one can estimate the statistical error through basic variance approximation techniques, while it is not straightforward to estimate the bias. The fact one can construct an unbiased simulation method at all is a remarkable result and relies on one being able to alter the driving SDE at the level of the PDE. That is, imagine we are interested in estimating $\mathbb{E}[G(X_T)]$, where the *driving* SDE is

$$dX_s = b(s, X_s)ds + \sigma(s, X_s)dW_s\,, \quad X_t = x\,.$$

Under sufficiently nice conditions one can view this as the solution to the linear terminal valued PDE

$$\begin{cases} \partial_t u(t,x) + b(t,x)\partial_x u(t,x) + \frac{\sigma(t,x)^2}{2}\partial_x^2 u(t,x) = 0\,, \\ u(T,x) = G(x)\,. \end{cases}$$

The idea now is to change the PDE to also change the driving SDE, namely we can equivalently consider the PDE

$$\begin{cases} \partial_t u(t,x) + b_0\partial_x u(t,x) + \frac{\sigma_0^2}{2}\partial_x^2 u(t,x) + (b(t,x)-b_0)\partial_x u(t,x) + \frac{\sigma(t,x)^2-\sigma_0^2}{2}\partial_x^2 u(t,x) = 0\,, \\ u(T,x) = G(x)\,. \end{cases}$$

The stochastic representation to this PDE is,

$$\mathbb{E}\left[G(\bar{X}_T) + \int_t^T (b(s,\bar{X}_s)-b_0)\partial_x u(s,\bar{X}_s) + \frac{\sigma(s,\bar{X}_s)^2-\sigma_0^2}{2}\partial_x^2 u(s,\bar{X}_s)\mathrm{d}s\right]\,,$$

with driving SDE,

$$\mathrm{d}\bar{X}_s = b_0\mathrm{d}s + \sigma_0\mathrm{d}W_s\,, \quad \bar{X}_t = x\,.$$

Provided we can simulate $\bar{X}$ exactly (which is the case for b_0 and σ_0 constant) then we can apply a branching diffusion algorithm to solve this PDE and hence the original problem. Crucially, since we simulate the SDE exactly, there is no error coming from the numerical scheme under which we simulate the SDE. We will not go into any further detail here as we use and explain these arguments and techniques in detail in the next chapter. Although we encourage the reader to see [HLTT17] and [DOW17] for details on the method outlined above.

Remark 12.2.1 (Simpler Notation)**.** *One nice feature of this method is that unlike the previous branching diffusion algorithms, here we only ever have one particle. Therefore we do not need to index the particle and this keeps the notation simpler. In Chapter 13 we shall also be in the situation whereby we only have one particle, hence we can also use simpler notation.*

Remark 12.2.2 (Alternative methods)**.** *There are alternative methods one can use to simulate an SDE exactly. One alternative is the so-called parametrix method which relies on measure changes, see* [AKH17]*. As it turns out, this is very similar to the representation one obtains via regime switching. Also* [RG15] *construct an unbiased estimator based on a sequence of approximations.*

However, one of the main advantages of branching diffusions for unbiased simulation is that it fits well in the framework of probabilistic PDE simulation, see [HLOT$^+$16]*. Therefore we can truly construct an unbiased representation of the PDE.*

Chapter 13

Representation for Transport PDEs

In this chapter we focus on transport PDEs, which are PDEs with no dependence on second spatial derivative, that is can be written as,

$$\begin{cases} \partial_t v(t,x) + b(t,x) \cdot Dv(t,x) = f(t,x,v,Dv)\,, \\ v(T,x) = g(x)\,. \end{cases}$$

Specifically we look to develop a stochastic representation for transport PDEs (under some assumptions). One of the main limitations when using Itô based stochastic techniques to represent PDEs is the requirement that the PDE is of second order in space (i.e. a "Laplacian" must be present). Indeed, the PDEs considered in Chapter 12 were all second order, thus PDEs with only one spatial and one time derivative (transport PDEs) have been, until now beyond the scope of stochastic techniques. One idea to navigate around this is to perturb the PDE by a "small" Laplacian, then one can use stochastic techniques on the perturbed PDE. Although this does provide a way to approximate the solution, it is very dependent on the perturbation being small enough so that the solution of the perturbed PDE is close to the first order PDE. Of course introducing a perturbation will lead to an error (bias) in the estimation, but more problematic is that the inverse of the perturbation coefficient will appear in the nonlinearities containing derivatives, thus the small perturbation makes the numerical scheme unstable. We discuss this point further in Section 13.4. Let us note that stochastic representations are only important for transport PDEs with nonlinearities in the derivative of the solution, see Remark 13.0.1.

As discussed in the previous chapter branching algorithms offer a useful approach to solve non-linear PDEs and also for unbiased simulation of SDEs (see [HLOT$^+$16], [DOW17]), via Monte Carlo methods. However, in order to apply Monte Carlo one requires estimators to be square-integrable and of finite computational complexity. For square integrability several works have fine tuned previous results to allow for increasing general cases: [HLTT17] introduced a control variate on the final step, which allowed for an unbiased simulation of an SDE with constant diffusion; later, [HLOT$^+$16] changed the time stepping scheme from an Exponential to a Gamma random variable, this allowed for the simulation of semilinear PDEs; most recently, [DOW17] used antithetic variables as well as control variates to obtain an unbiased algorithm for an SDE with non constant diffusion.

The material we present requires all of the above mentioned improvements along with new ideas in order to ensure the estimator to be square-integrable. Taking the long view, we believe these techniques to be crucial in extending this type of stochastic representations to the fully non-linear case. The second order parabolic fully nonlinear case has been considered in [War17] and [War18], but the theoretical basis for that case is to the best of our knowledge open. There are also several works looking at branching style algorithms but to tackle different types of PDEs, see [CT17], [AC17] and [HLT18] for further results.

Our Contributions. The contributions of this work are two-fold. Firstly we show how one can take the ideas of branching diffusions and regime switching to construct an unbiased stochastic representation for transport PDE. To the best of our knowledge this is the first result of its kind. Secondly, we improve upon the techniques currently presented in the literature [HLOT$^+$16], [DOW17], [HLTT17] in order to show our representation is square integrable and of finite computational complexity and thus can be used in Monte Carlo simulation. For better readability we also provide a heuristic description of our ideas.

From a *methodological* point of view, the approach presented is related to the regime switching algorithms presented in [DOW17] and [HLTT17], where one adds and subtracts terms in the PDE to change the "driving SDE" defined by the Dynkin operator. Such algorithms were inspired by branching diffusion algorithms as developed in [RRM10] and [HL12]. Here we add and subtract the second order derivative, which leaves us with a nonlinear PDE that can then be solved using regime switching (essentially we perturb the PDE then correct for the perturbation). Crucially this does not require σ to be small. Although the transport PDE we consider is simple, one of the main challenges is to keep the representation square integrable, which comes from the added second order term. The general case (fully nonlinear) is left as an open problem, nonetheless we give numerical examples showing that the general case is within (numerical) reach.

Remark 13.0.1. *Basic first order PDEs can easily be made to have a stochastic like representation using branching type arguments, for example a PDE of the type,*

$$\partial_t u(t,x) + b(t,x)\partial_x u(t,x) + u(t,x)^2 = 0, \quad u(T,x) = g(x)\,.$$

It is possible to write the solution to this as,

$$u(t,x) = g(X_T) + \int_t^T u(s,X_s)^2 \mathrm{d}s\,,$$

where X is the deterministic process satisfying the ODE $\mathrm{d}X_s = b(t,X_s)\mathrm{d}s$, $X_t = x$. Introducing random times into the solution of u as is done in standard branching we can obtain a solution to u as the expected product of particles at time T. A similar argument can also be made for nonlinear ODEs.

What is crucial here though is that this argument only holds when we do not have nonlinearities in the first derivative of the process, since we require Malliavin integration by parts tricks to deal with those. This is also the case when we want to apply the unbiased trick to b.

This work is organised as follows. In Section 13.1 we present our notation, the problem and give a heuristic description of our ideas. In Section 13.2 we present and prove our main results. Finally Section 13.4 illustrates numerically our findings to show our method is indeed unbiased. Moreover, we show the capability of our method to tackle problems in the nonlinear setting where the perturbation technique performs poorly.

13.1 Regime Switching Diffusion Representation

13.1.1 Notation and recap of stochastic representations

Consider a multidimensional stochastic differential equation (SDE) X starting at time point t, $0 \leq t \leq T$ of the form,

$$\mathrm{d}X_s = b(s,X_s)\mathrm{d}s + \sigma(s,X_s)\mathrm{d}W_s\,, \quad \text{for } s \in [t,T] \text{ and } X_t = x\,,$$

where the drift $b : [0,T] \times \mathbb{R}^d \to \mathbb{R}^d$ and diffusion $\sigma : [0,T] \times \mathbb{R}^d \to \mathbb{R}^{d\times d}$ satisfy the usual Lipschitz conditions so that the above SDE has a unique strong solution.

We associate with the SDE the infinitesimal generator $\mathcal{L}$, which when applied to any function $\phi \in C_b^{1,2}([0,T] \times \mathbb{R}^d, \mathbb{R})$ in the domain of $\mathcal{L}$ is,

$$(\mathcal{L}\phi)(t,x) = b(t,x) \cdot D\phi(t,x) + \frac{1}{2}a(t,x) : D^2\phi(t,x)\,, \quad \text{for all } (t,x) \in [0,T] \times \mathbb{R}^d\,,$$

where we define $a(t,x) = \sigma(t,x)\sigma(t,x)^\intercal$, $A : B := \text{trace}(AB^\intercal)$, $^\intercal$ is the transpose of a matrix and D, D^2 denotes the usual multi-dimensional spatial differential operators of order one and two (see [Eva98]).

It well known by the Feynman-Kac formula that if a unique classical solution $v \in C_b^{1,2}$ exists to the following PDE,

$$\begin{cases} \partial_t v(t,x) + \mathcal{L}v(t,x) = 0\,, \\ v(T,x) = g(x)\,, \end{cases}$$

for g a Lipschitz continuous function, then the solution of this PDE admits a stochastic representation, $v(t,x) = \mathbb{E}[g(X_T)|X_t = x]$.

13.1.2 Heuristic derivation of the idea of our work

To aid the presentation we give an introductory outline of our work. Although the ideas are inspired by those of branching diffusions and regime switching, how we construct the representation is different to that considered in Chapter 12. The ultimate goal here is to construct a stochastic representation of PDEs with only first order spatial derivatives and develop a way to deal with the corresponding 2nd order nonlinearity. We consider PDEs of the form

$$\begin{cases} \partial_t v(t,x) + b(t,x) \cdot Dv(t,x) = 0\,, \\ v(T,x) = g(x)\,, \end{cases} \tag{13.1.1}$$

for notational convenience we will work in one spatial dimension here (hence $D = \partial_x$). The problem with constructing a stochastic representation involving the use of Itô's formula is that we automatically obtain a second order derivative. However, it is known that arguments from branching diffusion can be used to deal with higher order derivatives through the Bismut-Elworthy-Li formula (automatic differentiation as developed in [FLL$^+$99]). Let us assume that v solving (13.1.1) is a unique classical solution which is $C_b^{1,2}$ (i.e. we can apply Itô's formula to v), then we can consider the following equivalent PDE

$$\begin{cases} \partial_t v(t,x) + b(t,x)\partial_x v(t,x) + \frac{1}{2}\sigma_0^2 \partial_{xx} v(t,x) - \frac{1}{2}\sigma_0^2 \partial_{xx} v(t,x) = 0\,, \\ v(T,x) = g(x)\,, \end{cases}$$

where σ_0 is some constant. In fact, as considered in [HLTT17], we can consider the equivalent PDE,

$$\begin{cases} \partial_t v(t,x) + b_0 \partial_x v(t,x) + \frac{1}{2}\sigma_0^2 \partial_{xx} v(t,x) + \big(b(t,x) - b_0\big)\partial_x v(t,x) - \frac{1}{2}\sigma_0^2 \partial_{xx} v(t,x) = 0\,, \\ v(T,x) = g(x)\,, \end{cases} \tag{13.1.2}$$

where b_0 is also some constant.

Stochastic Representation. Using the Feynman-Kac formula one can easily obtain the following stochastic representation of the solution to (13.1.2),

$$v(t,x) = \mathbb{E}\left[g(\bar{X}_T) + \int_t^T \big(b(s,\bar{X}_s) - b_0\big)\partial_x v(s,\bar{X}_s) - \frac{1}{2}\sigma_0^2 \partial_{xx} v(s,\bar{X}_s) \mathrm{d}s \;\Big|\; \bar{X}_t = x \right]\,, \tag{13.1.3}$$

where the driving SDE satisfies*

$$\mathrm{d}\bar{X}_s = b_0 \mathrm{d}s + \sigma_0 \mathrm{d}W_s\,, \quad \bar{X}_t = x \quad s \in [t,T]. \tag{13.1.4}$$

One can observe that such a representation holds provided our constants are $\mathcal{F}_t$ measurable.

Introduce a new random variable. Following a standard branching diffusion style argument, alongside the Brownian motion, W, we also consider an independent random variable τ with density $f > 0$ on $[0, T - t + \epsilon]$ for $\epsilon > 0$ and denote by $\overline{F}$ the corresponding survival function, namely for $s \in \mathbb{R}_+$ $\overline{F}(s) := \int_s^\infty f(r)\mathrm{d}r$. Consider some nice functions ψ and ϕ, then following

*Since we only have one "particle" in this set up we do not require the complex notation of Chapter 12. Hence we make the notation as simple as possible by dropping the superscripts on the diffusion process.

representation holds

$$\psi(T) + \int_t^T \phi(s)\mathrm{d}s = \mathbb{E}_f \left[\mathbb{1}_{\{\tau \geq T-t\}} \frac{\psi(T)}{\overline{F}(T-t)} + \mathbb{1}_{\{\tau < T-t\}} \frac{1}{f(\tau)} \phi(t+\tau) \right],$$

where $\mathbb{E}_f$ denotes the expectation for the random variable τ.

Rewriting the stochastic representation (13.1.3). Applying this to the Feynman-Kac representation (13.1.3) yields,

$$\begin{aligned} v(t,x) =& \mathbb{E}\Bigg[\frac{g(\bar{X}_T)}{\overline{F}(T-t)} \mathbb{1}_{\{t+\tau \geq T\}} + \mathbb{1}_{\{t+\tau < T\}} \frac{1}{f(\tau)} \Big[-\frac{1}{2}\sigma_0^2 \partial_{xx} v(t+\tau, \bar{X}_{t+\tau}) \\ & + \big(b(t+\tau, \bar{X}_{t+\tau}) - b_0\big) \partial_x v(t+\tau, \bar{X}_{t+\tau}) \Big] \,\Big|\, \bar{X}_t = x \Bigg]. \end{aligned} \quad (13.1.5)$$

One may note the abuse of notation here, the original Feynman-Kac representation expectation was only w.r.t. the Brownian motion, while (13.1.5) is w.r.t. both τ and the Brownian motion. To make the notation easier we now introduce the following stochastic sequence of times (stochastic mesh on the interval $[t,T]$), $t =: T_0 < T_1 < \cdots < T_{N_T} < T_{N_T+1} := T$ constructed as follows, take a sequence of i.i.d. copies of τ, then set $T_{k+1} = (T_k + \tau^{(k)}) \wedge T$ for $k \in \Lambda \subset \mathbb{N}$, where Λ is the set of integers (of stochastic length) $\{1, \dots, N_T + 1\}$. Using this mesh we then define $\Delta T_{k+1} = T_{k+1} - T_k$ and $\Delta W_{T_{k+1}} = W_{T_{k+1}} - W_{T_k}$.

Choosing the SDE's coefficients. Let us now consider a good choice of constant for b_0 (we define σ_0 later). As discussed in [HLTT17, DOW17], one can use the so called *frozen coefficient* function which defines the Euler scheme. That is, we may define the SDE $\bar{X}$ recursively over the random mesh by

$$\bar{X}_{T_k} = \bar{X}_{T_{k-1}} + b(T_{k-1}, \bar{X}_{T_{k-1}})\Delta T_k + \sigma_{k-1}\Delta W_{T_k}, \quad \bar{X}_0 = x, \quad (13.1.6)$$

for $k \in \Lambda$. Define θ_{k-1} as the times in the mesh and position of the SDE up to time T_{k-1} i.e. $\theta_{k-1} := (T_1, \dots, T_{k-1}, x, \bar{X}_{T_1}, \dots, \bar{X}_{T_{k-1}})$. Furthermore define the functions $\bar{b}(\theta_{k-1}, s, \bar{X}_s) = b(T_{k-1}, \bar{X}_{T_{k-1}})$ and $\sigma(\theta_{k-1}, s) = \sigma_{k-1}$ for $T_{k-1} < s$. Then the SDE defined recursively by,

$$\bar{X}_{T_k} = \bar{X}_{T_{k-1}} + \int_{T_{k-1}}^{T_k} \bar{b}(\theta_{k-1}, s, \bar{X}_s)\mathrm{d}s + \int_{T_{k-1}}^{T_k} \sigma(\theta_{k-1}, s)\mathrm{d}W_s, \quad (13.1.7)$$

is the Euler scheme in (13.1.6). Moreover, it is clear that the coefficients $\bar{b}(\theta_k, \cdot)$ and $\sigma(\theta_k, \cdot)$ are $\mathcal{F}_{T_k}$-adapted, hence can be used in (13.1.5). Using the coefficients coming from the Euler scheme is key here since we can simulate an Euler scheme exactly and hence the SDE appearing in (13.1.5) can be simulated exactly (which leads to the unbiased representation).

Remark 13.1.1. *We draw attention to a subtlety in the notation, we will define σ on intervals of the form $(\cdot, \cdot]$, thus σ is constant over each interval in the time mesh (as is the case in the Euler scheme).*

Obtaining a representation for the derivatives. The only terms left to consider in (13.1.5) are the derivatives of v. We will formulate rigorous results in Section 13.2, for now let us assume that all functions are sufficiently smooth and with good properties. We construct the Bismut-Elworthy-Li formula (automatic differentiation) w.r.t. the SDE (13.1.7). From [FLL$^+$99, Assumption 3.1] the following integration by parts relation holds for any square integrable function ϕ,

$$\partial_x \mathbb{E}[\phi(X_s) | X_t = x] = \mathbb{E}\left[\phi(X_s) \int_t^s \sigma(u)^{-1} Y(u) \mu(u) \mathrm{d}W_u \,\Big|\, X_t = x \right],$$

where Y is the first variation process of the SDE X and μ is any function such that $\int_t^s \mu(u)\mathrm{d}u = 1$. In the case of the SDE being (13.1.7), it is clear that the first variation process is constant equal to one (note σ does not have a space dependence). Typically one takes constant $\mu = 1/(s-t)$, thus for (13.1.7) we obtain,

$$\partial_x \mathbb{E}[\phi(X_{T_1})|X_t = x] = \mathbb{E}\left[\phi(X_{T_1})\frac{1}{\Delta T_1}\int_t^{T_1} \sigma(\theta_0, u)^{-1}\mathrm{d}W_u \;\Big|\; X_t = x\right] ,$$

The same method yields a similar expression for the second derivative

$$\partial_{xx} \mathbb{E}[\phi(X_{T_1})|X_t = x] = \mathbb{E}\left[\frac{\phi(X_{T_1})}{\Delta T_1^2}\left(\left(\int_t^{T_1} \sigma(\theta_0, u)^{-1}\mathrm{d}W_u\right)^2 - \int_t^{T_1}(\sigma(\theta_0, u)^{-1})^2\mathrm{d}u\right) \;\Big|\; X_t = x\right] ,$$

From this result and using the fact that σ is constant between mesh points we obtain for the second derivative

$$\begin{aligned}\partial_{xx}v(t,x) =&\mathbb{E}\Bigg[\frac{g(X_T)}{\overline{F}(\Delta T_1)}\mathbb{1}_{\{T_1\geq T\}}\frac{1}{\Delta T_1^2}\left(\left(\sigma(\theta_0,T_1)^{-1}\Delta W_{T_1}\right)^2 - (\sigma(\theta_0,T_1)^{-1})^2\Delta T_1\right)\\ &+\frac{\mathbb{1}_{\{T_1<T\}}}{f(\Delta T_1)}\left(\left(b(T_1,\bar{X}_{T_1}) - \bar{b}(\theta_0,T_1,\bar{X}_{T_1})\right)\partial_x v(T_2,\bar{X}_{T_2}) - \frac{1}{2}\sigma(\theta_0,T_1)^2\partial_{xx}v(T_1,X_{T_1})\right)\\ &\times\frac{\left(\sigma(\theta_0,T_1)^{-1}\Delta W_{T_1}\right)^2 - (\sigma(\theta_0,T_1)^{-1})^2\Delta T_1}{\Delta T_1^2}\;\Big|\; X_t = x\Bigg],\end{aligned}$$

the $\partial_x v$ term is similar. The idea of branching diffusion style algorithms is to continuously substitute in terms involving the solution until we remove the dependence on it. Of course, $v(t,x)$ does not appear inside the expectation, however, by using the tower property and flow property of the SDE we are able to derive the corresponding representations for $\partial_x v(T_k, X_{T_k})$ and $\partial_{xx}v(T_k, X_{T_k})$.

Rewriting the stochastic representation (13.1.5). Substituting in the expressions for $\partial_x v(T_1, X_{T_1})$ and $\partial_{xx}v(T_1, X_{T_1})$ into (13.1.5) yields,

$$\begin{aligned}&v(t,x)\\ &= \mathbb{E}\Bigg[\frac{g(\bar{X}_T)}{\overline{F}(\Delta T_1)}\mathbb{1}_{\{T_1\geq T\}} + \mathbb{1}_{\{T_1<T\}}\frac{1}{f(\Delta T_1)}\mathbb{E}\Bigg[\overline{\mathcal{W}}_2\Bigg\{\frac{g(\bar{X}_T)}{\overline{F}(\Delta T_2)}\mathbb{1}_{\{T_2\geq T\}} + \mathbb{1}_{\{T_2<T\}}\frac{1}{f(\Delta T_2)}\\ &\times\left[\left(b(T_2,\bar{X}_{T_2}) - \bar{b}(\theta_1,T_2,\bar{X}_{T_2})\right)\partial_x v(T_2,\bar{X}_{T_2}) - \frac{1}{2}\sigma(\theta_1,T_2)^2\partial_{xx}v(T_2,\bar{X}_{T_2})\right]\Bigg\}\;\Bigg|\;\bar{X}_{T_1}\Bigg]\;\Bigg|\;\bar{X}_t = x\Bigg],\end{aligned}$$

where $\overline{\mathcal{W}}_k$ is the so-called Malliavin weight stemming from the automatic differentiation,

$$\overline{\mathcal{W}}_k := \frac{b(T_{k-1},\bar{X}_{T_{k-1}}) - \bar{b}(\theta_{k-2},T_{k-1},\bar{X}_{T_{k-1}})}{\sigma(\theta_{k-1},T_k)}\frac{\Delta W_{T_k}}{\Delta T_k} - \frac{1}{2}\frac{\sigma(\theta_{k-2},T_{k-1})^2}{\sigma(\theta_{k-1},T_k)^2}\left(\frac{\Delta W_{T_k}^2 - \Delta T_k}{\Delta T_k^2}\right).$$

One observes that this Feynman-Kac representation now only depends on the solution v if $T_2 < T$.

Taking the limit. Following the standard procedure in branching diffusions (see [HL12, HLTT14, HLOT+16]), executing the same argument multiple times removes the dependence on v on the right hand side. Following [DOW17] we introduce the following notation,

$$M_{k+1} = \Delta b_k\sigma(\theta_k,T_{k+1})^{-1}\frac{\Delta W_{T_{k+1}}}{\Delta T_{k+1}} \quad\text{and}\quad V_{k+1} = -\frac{1}{2}\frac{\sigma(\theta_{k-1},T_k)^2}{\sigma(\theta_k,T_{k+1})^2}\left(\frac{\Delta W_{T_{k+1}}^2 - \Delta T_{k+1}}{\Delta T_{k+1}^2}\right). \tag{13.1.8}$$

where $\Delta b_k = b(T_k,\bar{X}_{T_k}) - \bar{b}(\theta_{k-1},T_k,\bar{X}_{T_k}) = b(T_k,\bar{X}_{T_k}) - b(T_{k-1},\bar{X}_{T_{k-1}})$. Further define the terms

$$P_{k+1} := \frac{M_{k+1} + \frac{1}{2}V_{k+1}}{f(\Delta T_k)} \quad \text{for } \ k \in \Lambda. \tag{13.1.9}$$

It is then clear that the solution to the PDE can be written as follows,

$$v(t,x) = \mathbb{E}\left[\frac{g(\bar{X}_{T_{N_T+1}})}{\bar{F}(\Delta T_{N_T+1})} \prod_{k=2}^{N_T+1} P_k \,\middle|\, \bar{X}_t = x \right] . \tag{13.1.10}$$

Although this relation is useful for us, in its current form it is not square integrable, thus we need to use some variance reduction techniques in order to use Monte Carlo. Moreover, many of the operations above require some form of integrability, these points will be the main focus of the next section.

13.2 Stochastic Representation

We look to derive a square-integrable representation that solves a PDE of the form,

$$\begin{cases} \partial_t v(t,x) + b(t) \cdot Dv(t,x) = 0 \quad \text{for all } (t,x) \in [0,T) \times \mathbb{R}^d \,, \\ v(T,x) = g(x) \,. \end{cases} \tag{13.2.1}$$

We wish to consider SDEs of the form (13.1.7), in d-dimensions this is,

$$\mathrm{d}\bar{X}_s = \bar{b}(\theta, s)\mathrm{d}s + \sigma(\theta, s)\mathbb{I}_d \mathrm{d}W_s \,, \quad \text{for } s \in [t,T] \text{ and } \bar{X}_t = x \,,$$

where $\mathbb{I}_d$ is the d-dimensional identity matrix. Unlike typical stochastic representations, σ is not fixed by the PDE, thus we have the freedom to choose σ. Although, the representation is somewhat independent of the precise choice of σ, the variance of the estimate (and hence the usefulness) heavily depends on σ.

Remark 13.2.1. *We show the representation in the case b is independent of space. Our original case was that detailed in Section 13.3.1, however, we were not able to obtain finite variance. We therefore make the assumption b only depends on time and return to the case of space dependency in Section 13.3.1, which is future work.*

In order to keep our representation and in particular our proofs as readable as possible, we consider only the one dimensional case. As one can clearly see though, due to fact that σ is a scalar multiplied by the identity, all our arguments generalise to the higher dimensional case.

The previous section outlined how one builds the stochastic representation without going into detail about when the various steps are applicable. We now want to show that this representation holds under some integrability and regularity assumptions. In the previous section we required two types of random variable, namely a driving Brownian motion and an i.i.d. sequence of random times $\tau^{(k)}$ with density f, independent of the Brownian motion and $k \in \Lambda$ as before. Thus consider the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ generated by these random variables, we also denote by $\mathbb{P}_W$ and $\mathbb{P}_f$ the probability measure ($\mathbb{E}_W$ and $\mathbb{E}_f$ the corresponding expectation) restricted to the Brownian motion and random times respectively. With this notation, one may think of $\mathbb{P}$ as the product measure $\mathbb{P}_W \otimes \mathbb{P}_f$. The corresponding filtration $\mathcal{F}_t$ is the sigma-algebra generated by the set of random times up to t i.e. $\max\{k : T_k \leq t\}$ and the Brownian motion up to t, hence, $\mathcal{F}_t := \sigma(T_1, \ldots, T_k, (W_s)_{s \leq t})$.

Let us first state the assumptions we will use.

Assumption 13.2.2. *We assume the drift, $b : [0,T] \to \mathbb{R}$ is uniformly Lipschitz in time.*

The analysis we carry out using regime switching techniques is sufficiently difficult to present that we assume the existence of a good enough solution to the transport PDE, as opposed to assuming sufficient conditions that would allow us to derive the said solution. Waiving the next assumption is left for future work.

Assumption 13.2.3. *Firstly we assume that there exists a unique solution $v \in C_b^{1,3}([0,T],\mathbb{R}^d)$ to* (13.2.1)*. In particular, we have that the terminal condition function g of the PDE satisfies $g \in C_b^2$.*

The assumption on g is not necessary since it follows from $v \in C_b^{1,3}$, however, we make this explicit since it is all we require for our estimator to be of finite variance. It is possible to put some conditions on b and g leading to a unique solution for general transport PDEs see [Kat75] for example. We do not go into detail here as this will again be the subject of future work.

We consider the particles to have a life time given by Gamma distributed random variables, i.e. τ has density,

$$f(s) := f_\Gamma^{\kappa,\eta}(s) = \frac{s^{\kappa-1}\exp(-s/\eta)}{\Gamma(\kappa)\eta^\kappa}\,, \quad \text{for all } \; s > 0 \; \text{ where } \; \kappa,\ \eta > 0\,, \tag{13.2.2}$$

where Γ is the Euler function $\Gamma(y) = \int_0^\infty x^{y-1}\exp(-x)\mathrm{d}x$.

We will use a *mesh dependent* coefficient for σ relying on the times at which the regime switching occurs,

$$\sigma(\theta_{k-1}, s) := \sigma_0 \prod_{i=1}^{k-1} \Delta T_i^n \quad \text{for } s \in (T_{k-1}, T_k]\,,\; k = 1,\dots,N_T+1\,,\;\; n \in \mathbb{R} \;\text{ and }\; \sigma_0 \in \mathbb{R}_+\,, \tag{13.2.3}$$

hence $\sigma(\theta_{k-1}, T_k) = \sigma_0 \prod_{i=1}^{k-1}\Delta T_i^n$, with the convention $\prod_{i=1}^0 \cdot = 1$.

Remark 13.2.4 (Adaptedness of σ)**.** *Even though our σ depends on the stochastic mesh, it is $\mathcal{F}_t$-adapted. This is of fundamental importance to show that the estimator in* (13.2.5) *solves the PDE* (13.2.1).

We make an assumption on the parameters of σ and f.

Assumption 13.2.5. *The power exponent n in the diffusion coefficient* (13.2.3) *satisfies $n \leq -1$. The shape parameter of the Gamma random variable,* (13.2.2)*, is $\kappa = 1/2$.*

Remark 13.2.6. *Under Assumption 13.2.5, σ is a positive function bounded from below away from zero. The bounds on n and κ are mainly for convenience in order for the proof of Proposition 13.2.8 to follow.*

As was alluded to in Section 13.1, (13.1.10) was not useful since it did not have finite second moment. To solve this problem we employ variance reduction techniques, namely antithetic variables and control variates. Consider the following auxiliary random variables, $\beta := (\beta_1 + \beta_2)/2$ with

$$\begin{cases} \beta_1 := \dfrac{g(\bar{X}_{T_{N_T+1}}) - g(\bar{X}_{T_{N_T}} + b(T_{N_T})\Delta T_{N_T+1})}{\overline{F}(\Delta T_{N_T+1})}\,\dfrac{M_{N_T+1} + \frac{1}{2}V_{N_T+1}}{f(\Delta T_{N_T})}\,, \\ \beta_2 := \dfrac{g(\hat{X}_{T_{N_T+1}}) - g(\bar{X}_{T_{N_T}} + b(T_{N_T})\Delta T_{N_T+1})}{\overline{F}(\Delta T_{N_T+1})}\,\dfrac{-M_{N_T+1} + \frac{1}{2}V_{N_T+1}}{f(\Delta T_{N_T})}\,, \end{cases} \tag{13.2.4}$$

where $\hat{X}$ is the antithetic of $\bar{X}$ i.e. the Euler scheme defined by, $\hat{X}_{T_k} = \bar{X}_{T_{k-1}} + b(T_{k-1})\Delta T_k - \sigma(\theta_{k-1}, T_k)\Delta W_{T_k}$ and V and M as defined in (13.1.8). It is straightforward to see that the additional g term is a control variate since its input is independent of Brownian motion $\Delta W_{T_{N_T+1}}$. One can further understand (β_1, β_2) as an antithetic pair.

We now state our main result of the chapter.

Theorem 13.2.7. *[Representation Solves the PDE] Let Assumptions 13.2.2, 13.2.3 and 13.2.5 hold, and let us denote by $\hat{v} : [0,T] \times \mathbb{R} \to \mathbb{R}$ the following function,*

$$\hat{v}(t,x) := \mathbb{E}\left[\beta \prod_{k=2}^{N_T} P_k \mathbb{1}_{\{N_T \geq 1\}} \,\middle|\, \sigma(\theta_0, t),\ X_t = x\right] + \mathbb{E}\left[\frac{g(\bar{X}_{T_1})}{\overline{F}(\Delta T_1)}\mathbb{1}_{\{N_T = 0\}} \,\middle|\, \sigma(\theta_0, t),\ X_t = x\right], \tag{13.2.5}$$

with $\{P_k\}_k$ as defined in (13.1.9)*. Then $\hat{v}$ solves the PDE* (13.2.1)*, namely $\hat{v} = v$ (hence $\hat{v}$ is an unbiased estimator of v). Moreover, the stochastic process generating $\hat{v}$ is square integrable and hence of finite variance.*

Outline of proof The proof of Theorem 13.2.7 requires several steps which we show in the following order.

1. Take $\tilde{v}$ in (13.2.5), which is the expected value of a stochastic process (estimator).
2. Show that the estimator is square integrable, Proposition 13.2.8.
3. Show that under enough integrability a stochastic representation to (13.2.1) exists when a $C_b^{1,3}([0,T],\mathbb{R})$ solution exists, Theorem 13.2.10.
4. Show that (13.2.5), satisfies the integrability conditions in Theorem 13.2.10 and thus solves (13.2.1), Theorem 13.2.13.

13.2.1 Variance analysis for a specific diffusion coefficient

Since our regime switching algorithm does not create new particles, our computational complexity for any Monte Carlo realisation is only $O(C(N_T+1))$, since $T<\infty$, it is clear we have finite computational complexity. We therefore only need to consider the variance of the estimator. We obtain the following.

Proposition 13.2.8. *Let Assumptions 13.2.2, 13.2.3 and 13.2.5 hold. Then the random variable appearing in* (13.2.5)*,*

$$\beta \prod_{k=2}^{N_T} P_k \mathbb{1}_{\{N_T \geq 1\}} + \frac{g(\bar{X}_{T_1})}{\overline{F}(\Delta T_1)} \mathbb{1}_{\{N_T=0\}}$$

has finite variance.

Although this proof is argued in a similar style to the proof of Proposition 4.1 in [DOW17], there are many subtle differences and we overall require a more refined analysis of the various terms to ensure our estimator has finite second moment. We point in particular to the "Interval splitting" argument in order to deal with instability in the last time point of the random mesh. This is essential to deal with the second order term that appears.

Proof. [Finite variance of the estimator]. Consider $\overline{\mathcal{F}}_k$ the sigma-algebra generated by the set of random times up to T_{k+1} and the Brownian motion up to T_k, hence[†], $\overline{\mathcal{F}}_k := \sigma(T_1,\dots,T_{k+1},(W_s)_{s\leq T\wedge T_k})$.

Throughout the proof, for ease of writing we suppress the condition in the expectation of the process starting at x at time t.

In order to show finite variance we only need to show finite second moment (the dominant term), further note that due to the indicators we obtain no cross term. Looking first at the second term of (13.2.5), by the bounds on the coefficients on the SDE and the Lipschitz property of g we have $\mathbb{E}[g(\bar{X}_{T_1})^2]<\infty$, and $\overline{F}(T-t)>0$, thus we have finite variance on the second term. For the first term in (13.2.5), we can rewrite the second moment as,

$$\mathbb{E}\left[\left(\beta\prod_{k=2}^{N_T}P_k\right)^2\mathbb{1}_{\{N_T\geq 1\}}\right] = \sum_{\ell=1}^{\infty}\mathbb{E}\left[\left(\beta\prod_{k=2}^{N_T}P_k\right)^2\middle| N_T=\ell\right]\times\mathbb{P}[N_T=\ell]\,.$$

In order to tackle this term we split the proof into several steps by bounding various quantities then combining them together to show the sum is bounded. We also note that we often work with conditional expectations, hence statements involving them are to be understood in the $\mathbb{P}$-a.s. sense.

Step 1: Bounding $\mathbb{E}[\beta^2|\overline{\mathcal{F}}_{N_T}, N_T=\ell]$, for β from (13.2.4). As is standard practice when we only care about showing an estimate to be finite we use C to denote some finite constant which

[†] One should note the small but critical distinction between $\mathcal{F}_t$ and $\overline{\mathcal{F}}_k$.

can change over inequalities but crucially can only depend on “known” constants such as T etc. By the tower property we can rewrite any term in the sum as,

$$\mathbb{E}\left[\left(\beta\prod_{k=2}^{N_T}P_k\right)^2\middle|N_T=\ell\right]=\mathbb{E}\left[\mathbb{E}[\beta^2|\overline{\mathcal{F}}_{N_T},N_T=\ell]\prod_{k=2}^{N_T}P_k^2\middle|N_T=\ell\right]\,.$$

Rewriting β with M_{N_T+1} and V_{N_T+1} as common factors then using Young’s inequality we obtain,

$$\begin{aligned}
&\mathbb{E}\big[\,\beta^2|\bar{\mathcal{F}}_{N_T},N_T=\ell\,\big]\\
&\leq C\mathbb{E}\left[\frac{\Big(g(\bar{X}_{T_{N_T+1}})-g(\hat{X}_{T_{N_T+1}})\Big)^2}{\overline{F}(\Delta T_{N_T+1})^2}\frac{M_{N_T+1}^2}{f(\Delta T_{N_T})^2}\middle|\overline{\mathcal{F}}_{N_T},N_T=\ell\right]\\
&+C\mathbb{E}\left[\frac{\Big(g(\bar{X}_{T_{N_T+1}})+g(\hat{X}_{N_T+1})-2g(\bar{X}_{T_{N_T}}+b(T_{N_T})\Delta T_{N_T+1})\Big)^2}{\overline{F}(\Delta T_{N_T+1})^2}\frac{\frac{1}{2}V_{N_T+1}^2}{f(\Delta T_{N_T})^2}\middle|\overline{\mathcal{F}}_{N_T},N_T=\ell\right]\,.
\end{aligned}$$

Considering the first term on the RHS, we note by the Lipschitz property of g that,

$$|g(\bar{X}_{T_{N_T+1}})-g(\hat{X}_{T_{N_T+1}})|\leq L|\bar{X}_{T_{N_T+1}}-\hat{X}_{T_{N_T+1}}|\leq C|\sigma(\theta_{N_T},T_{N_T+1})\Delta W_{T_{N_T+1}}|\,.$$

Hence using this bound and the representation for M_{N_T+1} (see (13.1.8)),

$$\begin{aligned}
&\mathbb{E}\left[\frac{\Big(g(\bar{X}_{T_{N_T+1}})-g(\hat{X}_{T_{N_T+1}})\Big)^2}{\overline{F}(\Delta T_{N_T+1})^2}\frac{M_{N_T+1}^2}{f(\Delta T_{N_T})^2}\middle|\overline{\mathcal{F}}_{N_T},N_T=\ell\right]\\
&\leq C\frac{\Delta b_{N_T}^2}{f(\Delta T_{N_T})^2}\mathbb{E}\left[\Big(\Delta W_{T_{N_T+1}}\sigma(\theta_{N_T},T_{N_T+1})\Big)^2\left(\frac{\Delta W_{T_{N_T+1}}}{\Delta T_{N_T+1}}\sigma(\theta_{N_T},T_{N_T+1})^{-1}\right)^2\middle|\overline{\mathcal{F}}_{N_T},N_T=\ell\right]\\
&=C\frac{\Delta b_{N_T}^2}{f(\Delta T_{N_T})^2}\,,
\end{aligned}$$

where we used $1/\overline{F}(\Delta T_{N_T+1})^2\leq C$ in the inequality. For the second term on the RHS, it is more complex, let us first split the terms using Cauchy-Schwarz,

$$\begin{aligned}
&\mathbb{E}\left[\Big(g(\bar{X}_{T_{N_T+1}})+g(\hat{X}_{N_T+1})-2g(\bar{X}_{T_{N_T}}+b(T_{N_T})\Delta T_{N_T+1})\Big)^2V_{N_T+1}^2\middle|\overline{\mathcal{F}}_{N_T},N_T=\ell\right]\\
&\leq\mathbb{E}\left[\Big(g(\bar{X}_{T_{N_T+1}})+g(\hat{X}_{N_T+1})-2g(\bar{X}_{T_{N_T}}+b(T_{N_T})\Delta T_{N_T+1})\Big)^4\middle|\overline{\mathcal{F}}_{N_T},N_T=\ell\right]^{1/2}\\
&\quad\times\mathbb{E}\left[V_{N_T+1}^4\middle|\overline{\mathcal{F}}_{N_T},N_T=\ell\right]^{1/2}\,.
\end{aligned}$$

Let us firstly focus on the g term. Consider the ODE on the interval $s\in[T_{N_T},T_{N_T+1}]$,

$$\frac{\mathrm{d}Y_s}{\mathrm{d}s}=b(T_{N_T})\,,\qquad Y_{T_{N_T}}=\bar{X}_{T_{N_T}}\,.$$

Then, the solution is $Y_{T_{N_T+1}}=\bar{X}_{T_{N_T}}+b(T_{N_T})\Delta T_{N_T+1}$. Consequently,

$$\begin{aligned}
&g\Big(\bar{X}_{T_{N_T}}+b(T_{N_T})\Delta T_{N_T+1}\Big)-g(\bar{X}_{T_{N_T}})\\
&\quad=\int_{T_{N_T}}^{T_{N_T+1}}g'(Y_s)\mathrm{d}Y_s=\int_{T_{N_T}}^{T_{N_T+1}}g'\Big(\bar{X}_{T_{N_T}}+b(T_{N_T})(s-T_{N_T})\Big)b\big(T_{N_T}\big)\mathrm{d}s\,.
\end{aligned}\tag{13.2.6}$$

By applying Itô's formula to $g(\bar{X}_{T_{N_T+1}})$ and $g(\hat{X}_{T_{N_T+1}})$ (recall $g \in C_b^2$), and using (13.2.6) we obtain,

$$
\begin{aligned}
&g(\bar{X}_{T_{N_T+1}}) + g(\hat{X}_{T_{N_T+1}}) - 2g(\bar{X}_{T_{N_T}} + b(T_{N_T})\Delta T_{N_T+1}) \\
&= \frac{1}{2}\sigma(\theta_{N_T}, T_{N_T+1})^2 \int_{T_{N_T}}^{T_{N_T+1}} (g''(\bar{X}_s) + g''(\hat{X}_s))\mathrm{d}s + \sigma(\theta_{N_T}, T_{N_T+1}) \int_{T_{N_T}}^{T_{N_T+1}} (g'(\bar{X}_s) - g'(\hat{X}_s))\mathrm{d}W_s \\
&\quad + \int_{T_{N_T}}^{T_{N_T+1}} \Big(g'(\bar{X}_s) + g'(\hat{X}_s) - 2g'\Big(\bar{X}_{T_{N_T}} + b(T_{N_T})(s - T_{N_T})\Big)\Big) b(T_{N_T})\mathrm{d}s\,. \qquad (13.2.7)
\end{aligned}
$$

Since g' is Lipschitz, we obtain,

$$
\begin{aligned}
|g'(\bar{X}_s) - g'\Big(\bar{X}_{T_{N_T}} + b(T_{N_T})(s - T_{N_T})\Big)| &\le C|\bar{X}_s - \bar{X}_{T_{N_T}} + b(T_{N_T})(s - T_{N_T})| \\
&\le C\sigma(\theta_{N_T}, T_{N_T+1})|W_s - W_{T_{N_T}}|\,,
\end{aligned}
$$

the same bound holds for the $g(\hat{X}_s)$ term. Thus the following bound can be obtained for the final integral in (13.2.7)

$$
\begin{aligned}
&\int_{T_{N_T}}^{T_{N_T+1}} \Big(g'(\bar{X}_s) + g'(\hat{X}_s) - 2g'\Big(\bar{X}_{T_{N_T}} + b(T_{N_T})(s - T_{N_T})\Big)\Big) b(T_{N_T})\mathrm{d}s \\
&\qquad \le C|b(T_{N_T})|\sigma(\theta_{N_T}, T_{N_T+1}) \int_{T_{N_T}}^{T_{N_T+1}} |W_s - W_{T_{N_T}}|\mathrm{d}s\,.
\end{aligned}
$$

Recalling that we are interested in the fourth moment, using Doob's maximal inequality,

$$
\begin{aligned}
&\mathbb{E}\left[\left(\int_{T_{N_T}}^{T_{N_T+1}} |W_s - W_{T_{N_T}}|\mathrm{d}s\right)^4 \;\Big|\; \overline{\mathcal{F}}_{N_T}, N_T = \ell\right] \\
&\le C\Delta T_{N_T+1}^4 \mathbb{E}\left[\sup_{T_{N_T} \le s \le T_{N_T+1}} |W_s - W_{T_{N_T}}|^4 \;\Big|\; \overline{\mathcal{F}}_{N_T}, N_T = \ell\right] \le C\Delta T_{N_T+1}^6\,.
\end{aligned}
$$

For the stochastic integral in (13.2.7), again taking the fourth moment we obtain,

$$
\begin{aligned}
&\sigma(\theta_{N_T}, T_{N_T+1})^4 \mathbb{E}\left[\left(\int_{T_{N_T}}^{T_{N_T+1}} (g'(\bar{X}_s) - g'(\hat{X}_s))\mathrm{d}W_s\right)^4 \;\Big|\; \overline{\mathcal{F}}_{N_T}, N_T = \ell\right] \\
&\qquad = 3\sigma(\theta_{N_T}, T_{N_T+1})^4 \mathbb{E}\left[\left(\int_{T_{N_T}}^{T_{N_T+1}} (g'(\bar{X}_s) - g'(\hat{X}_s))^2\mathrm{d}s\right)^2 \;\Big|\; \overline{\mathcal{F}}_{N_T}, N_T = \ell\right]\,.
\end{aligned}
$$

Using that g' is Lipschitz and the difference is given by

$$
|g'(\bar{X}_s) - g'(\hat{X}_s)| \le C|\sigma(\theta_{N_T}, T_{N_T+1})(W_s - W_{T_{N_T}}) + \sigma(\theta_{N_T}, T_{N_T+1})(W_s - W_{T_{N_T}})|\,.
$$

This along with a similar Doob's maximal inequality implies that we can bound the stochastic integral by,

$$
\sigma(\theta_{N_T}, T_{N_T+1})^4 \mathbb{E}\left[\left(\int_{T_{N_T}}^{T_{N_T+1}} (g'(\bar{X}_s) - g'(\hat{X}_s))\mathrm{d}W_s\right)^4 \;\Big|\; \overline{\mathcal{F}}_{N_T}, N_T = \ell\right] \le C\sigma(\theta_{N_T}, T_{N_T+1})^8 \Delta T_{N_T+1}^4\,.
$$

Recalling that g'' is bounded, we can bound the remaining term in (13.2.7) by a similar term to the stochastic integral to obtain,

$$\mathbb{E}\Big[(g(\bar{X}_{T_{N_T+1}})+g(\hat{X}_{T_{N_T+1}})-2g(\bar{X}_{T_{N_T}}+b(T_{N_T})\Delta T_{N_T+1}))^4|\overline{\mathcal{F}}_{N_T},N_T=\ell\Big] \tag{13.2.8}$$
$$\leq C\sigma(\theta_{N_T},T_{N_T+1})^8\Delta T_{N_T+1}^4\,.$$

The above bound was obtained using differentiability and Itô's formula, however, it will also be useful for us to note that just using the Lipschitz property yields,

$$\mathbb{E}\Big[(g(\bar{X}_{T_{N_T+1}})+g(\hat{X}_{T_{N_T+1}})-2g(\bar{X}_{T_{N_T}}+b(T_{N_T})\Delta T_{N_T+1}))^4|\overline{\mathcal{F}}_{N_T},N_T=\ell\Big]$$
$$\leq C\sigma(\theta_{N_T},T_{N_T+1})^4\Delta T_{N_T+1}^2\,.$$

Hence we obtain the following stronger bound for the g terms

$$\mathbb{E}\Big[(g(\bar{X}_{T_{N_T+1}})+g(\hat{X}_{T_{N_T+1}})-2g(\bar{X}_{T_{N_T}}+b(T_{N_T})\Delta T_{N_T+1}))^4|\overline{\mathcal{F}}_{N_T},N_T=\ell\Big]$$
$$\leq C\min\left[\sigma(\theta_{N_T},T_{N_T+1})^4\Delta T_{N_T+1}^2,\sigma(\theta_{N_T},T_{N_T+1})^8\Delta T_{N_T+1}^4\right]\,.$$

For the V term,

$$\begin{aligned}\mathbb{E}\left[V_{N_T+1}^4|\overline{\mathcal{F}}_{N_T},N_T=\ell\right]&\leq C\frac{\sigma(\theta_{N_T-1},T_{N_T})^8}{\sigma(\theta_{N_T},T_{N_T+1})^8}\frac{1}{\Delta T_{N_T+1}^8}\mathbb{E}\left[\left(\Delta W_{T_{N_T+1}}^2-\Delta T_{N_T+1}\right)^4\Big|\overline{\mathcal{F}}_{N_T},N_T=\ell\right]\\&\leq C\frac{\sigma(\theta_{N_T-1},T_{N_T})^8}{\sigma(\theta_{N_T},T_{N_T+1})^8}\frac{1}{\Delta T_{N_T+1}^4}\,.\end{aligned}$$

Hence using Cauchy-Schwarz we obtain,

$$\mathbb{E}\left[\frac{\left(g(\bar{X}_{T_{N_T+1}})+g(\hat{X}_{T_{N_T+1}})-2g(\bar{X}_{T_{N_T}}+b(T_{N_T})\Delta T_{N_T+1})\right)^2}{\overline{F}(\Delta T_{N_T+1})^2}\frac{\frac{1}{2}V_{N_T+1}^2}{f(\Delta T_{N_T})^2}\Bigg|\overline{\mathcal{F}}_{N_T},N_T=\ell\right]$$
$$\leq C\frac{\sigma(\theta_{N_T-1},T_{N_T})^4}{\sigma(\theta_{N_T},T_{N_T+1})^4}\frac{1}{\Delta T_{N_T+1}^2}\min\left[\sigma(\theta_{N_T},T_{N_T+1})^2\Delta T_{N_T+1},\sigma(\theta_{N_T},T_{N_T+1})^4\Delta T_{N_T+1}^2\right]\frac{1}{f(\Delta T_{N_T})^2}\,.$$

Therefore, the conditional expectation of β^2 can be bounded by,

$$\mathbb{E}[\beta^2|\overline{\mathcal{F}}_{N_T},N_T=\ell]\leq\frac{C}{f(\Delta T_{N_T})^2}\left(\Delta b_{N_T}^2+\frac{\sigma(\theta_{N_T-1},T_{N_T})^4}{\sigma(\theta_{N_T},T_{N_T+1})^2}\frac{\min\left[1,\sigma(\theta_{N_T},T_{N_T+1})^2\Delta T_{N_T+1}\right]}{\Delta T_{N_T+1}}\right)\,.$$

Step 2: Bounding $\mathbb{E}[P_{k+1}^4|\overline{\mathcal{F}}_k,N_T=\ell]$. Let $k\in\Lambda$ and note by Assumption 13.2.2 we obtain,

$$\mathbb{E}\big[\Delta b_k^4|\overline{\mathcal{F}}_{k-1},N_T=\ell\big]\leq C\Delta T_k^4\,.$$

From (13.1.8) we observe the following,

$$\begin{aligned}\mathbb{E}[M_{k+1}^4|\overline{\mathcal{F}}_k,N_T=\ell]&\leq C\frac{\Delta b_k^4}{\Delta T_{k+1}^2}\frac{1}{\sigma(\theta_k,T_{k+1})^4}\leq C\frac{\Delta T_k^4}{\Delta T_{k+1}^2}\frac{1}{\sigma(\theta_k,T_{k+1})^4}\,,\\\mathbb{E}[V_{k+1}^4|\overline{\mathcal{F}}_k,N_T=\ell]&\leq C\frac{\sigma(\theta_{k-1},T_k)^8}{\sigma(\theta_k,T_{k+1})^8}\frac{1}{\Delta T_{k+1}^4}\,.\end{aligned}$$

By Assumption 13.2.5 and the fact that σ is bounded from below implies that the V term dominates the M term, hence, we obtain,

$$\mathbb{E}[P_{k+1}^4|\overline{\mathcal{F}}_k,N_T=\ell]\leq C\frac{1}{f(\Delta T_k)^4}\frac{\sigma(\theta_{k-1},T_k)^8}{\sigma(\theta_k,T_{k+1})^8}\frac{1}{\Delta T_{k+1}^4}. \tag{13.2.9}$$

We are now able to consider bounding the term we originally set out to. Using the bound we obtained for β^2,

$$
\begin{aligned}
&\mathbb{E}\left[\beta^2 \prod_{k=2}^{N_T} P_k^2 \middle| N_T = \ell\right] \\
&\leq \mathbb{E}\left[\frac{C}{f(\Delta T_{N_T})^2}\Big(\Delta b_{N_T}^2 + \frac{\sigma(\theta_{N_T-1}, T_{N_T})^4}{\sigma(\theta_{N_T}, T_{N_T+1})^2}\frac{\min\left[1, \sigma(\theta_{N_T}, T_{N_T+1})^2 \Delta T_{N_T+1}\right]}{\Delta T_{N_T+1}}\Big) \prod_{k=2}^{N_T} P_k^2 \middle| N_T = \ell\right].
\end{aligned}
\tag{13.2.10}
$$

One can view this product as having two components, one which does not depend on ΔT_{N_T+1} which comes from the Δb_{N_T} and a component that does depend on ΔT_{N_T+1}. In order to show that the second moment is finite we split these two components and show each of them is finite.

Step 3: Bounding each product in (13.2.10). Let us start by considering the product from the Δb_{N_T} term

$$
\mathbb{E}\left[\frac{\Delta b_{N_T}^2}{f(\Delta T_{N_T})^2} \prod_{k=2}^{N_T} P_k^2 \middle| N_T = \ell\right] = \mathbb{E}\left[\frac{1}{f(\Delta T_{N_T})^2}\mathbb{E}[\Delta b_{N_T}^2 P_{N_T}^2 | \mathcal{F}_{N_T-1}, N_T = \ell] \prod_{k=2}^{N_T-1} P_k^2 \middle| N_T = \ell\right].
$$

Applying Cauchy-Schwarz to the internal expectation and using the previous bounds we obtain,

$$
\mathbb{E}[\Delta b_{N_T}^2 P_{N_T}^2 | \mathcal{F}_{N_T-1}, N_T = \ell] \leq C \frac{1}{f(\Delta T_{N_T-1})^2} \frac{\sigma(\theta_{N_T-2}, T_{N_T-1})^4}{\sigma(\theta_{N_T-1}, T_{N_T})^4}. \tag{13.2.11}
$$

Note that this bound and (13.2.9) have no dependence on the Brownian motion, therefore we can isolate each P_k by recursively conditioning, i.e.

$$
\begin{aligned}
&\mathbb{E}\left[\frac{C}{f(\Delta T_{N_T})^2}\mathbb{E}[\Delta b_{N_T}^2 P_{N_T}^2 | \mathcal{F}_{N_T-1}, N_T = \ell] \prod_{k=2}^{N_T-1} P_k^2 \middle| N_T = \ell\right] \\
&\leq \mathbb{E}\left[\frac{C}{f(\Delta T_{N_T})^2} \frac{1}{f(\Delta T_{N_T-1})^2} \frac{\sigma(\theta_{N_T-2}, T_{N_T-1})^4}{\sigma(\theta_{N_T-1}, T_{N_T})^4} \prod_{k=2}^{N_T-1} P_k^2 \middle| N_T = \ell\right] \\
&= \mathbb{E}\left[\frac{C}{f(\Delta T_{N_T})^2} \frac{1}{f(\Delta T_{N_T-1})^2} \frac{\sigma(\theta_{N_T-2}, T_{N_T-1})^4}{\sigma(\theta_{N_T-1}, T_{N_T})^4}\mathbb{E}[P_{N_T-1}^2 | \mathcal{F}_{N_T-2}, \Delta T_{N_T}, N_T = \ell] \prod_{k=2}^{N_T-2} P_k^2 \middle| N_T = \ell\right].
\end{aligned}
$$

Using our results and noting that most of the σ terms cancel yields the following bound,

$$
\begin{aligned}
&\mathbb{E}\left[\frac{\Delta b_{N_T}^2}{f(\Delta T_{N_T})^2} \prod_{k=2}^{N_T} P_k^2 \middle| N_T = \ell\right] \\
&\leq \mathbb{E}\left[\frac{C^{N_T}}{f(\Delta T_{N_T})^2} \frac{1}{f(\Delta T_{N_T-1})^2} \frac{\sigma(\theta_{N_T-2}, T_{N_T-1})^4}{\sigma(\theta_{N_T-1}, T_{N_T})^4} \prod_{k=2}^{N_T-1} \frac{1}{f(\Delta T_{k-1})^2} \frac{\sigma(\theta_{k-2}, T_{k-1})^4}{\sigma(\theta_{k-1}, T_k)^4} \frac{1}{\Delta T_k^2} \middle| N_T = \ell\right] \\
&= \mathbb{E}\left[\frac{C^{N_T}}{f(\Delta T_{N_T})^2} \frac{1}{f(\Delta T_1)^2} \frac{\sigma(\theta_0, T_1)^4}{\sigma(\theta_{N_T-1}, T_{N_T})^4} \prod_{k=2}^{N_T-1} \frac{1}{f(\Delta T_k)^2} \frac{1}{\Delta T_k^2} \middle| N_T = \ell\right].
\end{aligned}
$$

Recall the goal here is to ultimately bound this by a term of the form C^{N_T}, which holds provided all ΔT_k dependence is to a positive power. Recall that since f is the density for the Gamma distribution with shape κ, we have that,

$$
f(\Delta T_k) \geq C \Delta T_k^{\kappa-1} \quad \implies \quad \frac{1}{f(\Delta T_{N_T})^2} \leq C \Delta T_{N_T}^{2-2\kappa}.
$$

Using the representation for σ we obtain terms of the form $\Delta T_k^{2-2\kappa-2-4n}$, hence we require $2\kappa - 4n \geq 0$, which suggests $n \leq -\kappa/2$. Since Assumption 13.2.5 implies these conditions on n

and κ hold[‡], one obtains

$$\mathbb{E}\left[\frac{\Delta b_{N_T}^2}{f(\Delta T_{N_T})^2}\prod_{k=2}^{N_T} P_k^2 \middle| N_T=\ell\right] \leq \mathbb{E}\left[C^{N_T}\middle| N_T=\ell\right]. \tag{13.2.12}$$

Showing this is finite is done in [DOW17]. As it turns out the other term in (13.2.10) also dominates this term, hence we do not discuss it further.

For the second term in (13.2.10) we note that the σ terms do not depend on the Brownian motion, hence we can again condition to isolate the various P_k terms, hence,

$$\begin{aligned}
&\mathbb{E}\left[\frac{C}{f(\Delta T_{N_T})^2}\frac{\sigma(\theta_{N_T-1},T_{N_T})^4}{\sigma(\theta_{N_T},T_{N_T+1})^2}\frac{\min\left[1,\sigma(\theta_{N_T},T_{N_T+1})^2\Delta T_{N_T+1}\right]}{\Delta T_{N_T+1}}\prod_{k=2}^{N_T} P_k^2\middle| N_T=\ell\right]\\
&\leq \mathbb{E}\left[\frac{C}{f(\Delta T_{N_T})^2}\frac{\sigma(\theta_{N_T-1},T_{N_T})^4}{\sigma(\theta_{N_T},T_{N_T+1})^2}\frac{\min\left[1,\sigma(\theta_{N_T},T_{N_T+1})^2\Delta T_{N_T+1}\right]}{\Delta T_{N_T+1}}\right.\\
&\qquad\left.\times\prod_{k=2}^{N_T}\frac{1}{f(\Delta T_{k-1})^2}\frac{\sigma(\theta_{k-2},T_{k-1})^4}{\sigma(\theta_{k-1},T_k)^4}\frac{1}{\Delta T_k^2}\middle| N_T=\ell\right].
\end{aligned} \tag{13.2.13}$$

By cancelling repeating σ terms in the product and again using $1/f(\Delta T_1)\leq C$, we obtain the following simpler result,

$$(13.2.13)\leq C\mathbb{E}\left[\frac{\sigma(\theta_0,T_1)^4}{\sigma(\theta_{N_T},T_{N_T+1})^2}\frac{\min\left[1,\sigma(\theta_{N_T},T_{N_T+1})^2\Delta T_{N_T+1}\right]}{\Delta T_{N_T+1}}\prod_{k=2}^{N_T}\frac{1}{f(\Delta T_k)^2}\frac{1}{\Delta T_k^2}\middle| N_T=\ell\right]. \tag{13.2.14}$$

Using the fact that $\sigma(\theta_0,T_1)=\sigma_0$ and f is the density for the Gamma distribution we can bound (13.2.14) by,

$$\begin{aligned}
&\mathbb{E}\left[C^{N_T}\frac{\min\left[1,\sigma(\theta_{N_T},T_{N_T+1})^2\Delta T_{N_T+1}\right]}{\sigma(\theta_{N_T},T_{N_T+1})^2}\Delta T_{N_T+1}^{-1}\prod_{k=2}^{N_T}\Delta T_k^{-2\kappa}\middle| N_T=\ell\right]\\
&\leq \mathbb{E}\left[C^{N_T}\frac{\sigma(\theta_{N_T},T_{N_T+1})^{\nu}\Delta T_{N_T+1}^{\nu/2}}{\sigma(\theta_{N_T},T_{N_T+1})^2}\Delta T_{N_T+1}^{-1}\prod_{k=2}^{N_T}\Delta T_k^{-2\kappa}\middle| N_T=\ell\right] \quad \text{for } \nu\in[0,2],
\end{aligned} \tag{13.2.15}$$

where the inequality comes from the observation that,

$$\min\left[1,\sigma(\theta_{N_T},T_{N_T+1})^2\Delta T_{N_T+1}\right]\leq\sigma(\theta_{N_T},T_{N_T+1})^{\nu}\Delta T_{N_T+1}^{\nu/2}\quad\text{for any } \nu\in[0,2].$$

The presence of $\Delta T_{N_T+1}^{-1}$ makes (13.2.15) more challenging. Of course, one could take $\nu=2$ to remove $\Delta T_{N_T+1}^{-1}$, however, this also removes σ and since $\kappa>0$ we are still left with an unbounded product. Therefore we must chose ν carefully and apply a delicate argument to appropriately bound (13.2.15).

One can note the similarity between (13.2.15) and (13.2.12). However, (13.2.15) is more complex and as it turns out, the bound we eventually achieve for it dominates (13.2.12). We therefore complete the proof showing (13.2.15) is bounded, since this implies (13.2.12) is bounded.

Step 4: Interval splitting. Recall we are interested in proving convergence of the sum

$$\sum_{\ell=1}^{\infty}\mathbb{E}\left[\left(\beta\prod_{k=2}^{\ell}P_k\right)^2\middle| N_T=\ell\right]\mathbb{P}[N_T=\ell].$$

Let us split this into two components, $\ell=1$ and $\ell\geq 2$. When $\ell=1$ we obtain nothing from the product and are thus only showing that β is square integrable, such is obvious from our previous

[‡] Note that $\kappa=1/2$ also implies $1/f(\Delta T_1)\leq C$.

calculations. We now concentrate on the case $\ell \geq 2$. Recall that for $i = 1, \dots, M$, if $Y_i \sim \Gamma(a, b)$ i.i.d. then $\sum_{i=1}^{M} Y_i \sim \Gamma(aM, b)$ and fix $\ell \geq 2$, we can then partition the expectation as follows,

$$\mathbb{E}\left[\left(\beta \prod_{k=2}^{\ell} P_k\right)^2 \middle| N_T = \ell\right] = \mathbb{E}\left[\left(\beta \prod_{k=2}^{\ell} P_k\right)^2 \middle| N_T = \ell, \Delta T_{N_T+1} \geq \frac{T}{\ell}\right] \mathbb{P}\left[\Delta T_{N_T+1} \geq \frac{T}{\ell} \middle| N_T = \ell\right]$$
$$+ \sum_{m=1}^{\infty} \mathbb{E}\left[\left(\beta \prod_{k=2}^{\ell} P_k\right)^2 \middle| N_T = \ell, \frac{T}{\ell^{m+1}} \leq \Delta T_{N_T+1} < \frac{T}{\ell^m}\right] \mathbb{P}\left[\frac{T}{\ell^{m+1}} \leq \Delta T_{N_T+1} < \frac{T}{\ell^m} \middle| N_T = \ell\right] .$$

Firstly, we note that when $\Delta T_{N_T+1} \geq T/\ell$, the expectation is simple to bound since we can take the minimum as 1 (the $\nu = 0$ case in (13.2.15)) then use the fact $\sigma(\theta_{N_T}, T_{N_T+1})^{-2} = \sigma_0^{-2} \prod_{i=1}^{\ell} \Delta T_i^{2n}$ and $\kappa < -n$ by Assumption 13.2.5. Hence the following bound holds,

$$\mathbb{E}\left[\left(\beta \prod_{k=2}^{\ell} P_k\right)^2 \middle| N_T = \ell, \Delta T_{N_T+1} \geq \frac{T}{\ell}\right] \mathbb{P}\left[\Delta T_{N_T+1} \geq \frac{T}{\ell} \middle| N_T = \ell\right] \leq \ell C^{\ell} .$$

For the case $m \geq 1$, we have that

$$\mathbb{P}\left[\frac{T}{\ell^{m+1}} \leq \Delta T_{N_T+1} < \frac{T}{\ell^m} \middle| N_T = \ell\right] = \mathbb{P}\left[T - \frac{T}{\ell^m} \leq \sum_{i=1}^{\ell} \Delta T_i < T - \frac{T}{\ell^{m+1}} \middle| N_T = \ell\right] .$$

Due to the fact $\kappa = 1/2$ by Assumption 13.2.5, the distribution of $\sum_{i=1}^{\ell} \Delta T_i$ is Gamma with shape parameter at least 1, therefore the density has a finite maximum, unfortunately the conditioning makes this probability difficult to deal with. We therefore expand,

$$\begin{aligned}
&\mathbb{P}\left[T - \frac{T}{\ell^m} \leq \sum_{i=1}^{\ell} \Delta T_i < T - \frac{T}{\ell^{m+1}} \middle| N_T = \ell\right] \\
&\quad = \frac{1}{\mathbb{P}[N_T = \ell]} \mathbb{P}\left[T - \frac{T}{\ell^m} \leq \sum_{i=1}^{\ell} \Delta T_i < T - \frac{T}{\ell^{m+1}},\ \sum_{i=1}^{\ell} \Delta T_i < T,\ \sum_{i=1}^{\ell+1} \Delta T_i \geq T\right] \\
&\quad \leq \frac{1}{\mathbb{P}[N_T = \ell]} \mathbb{P}\left[T - \frac{T}{\ell^m} \leq \sum_{i=1}^{\ell} \Delta T_i < T - \frac{T}{\ell^{m+1}}\right] .
\end{aligned}$$

Using this form we have removed the conditional dependence on the number of jumps and therefore we can use the distribution of $\sum_{i=1}^{\ell} \Delta T_i$. We note that for ℓ large the density of the distribution at point T will be larger than values less than T, further, since the density has a finite maximum, for ℓ smaller we can bound by some constant multiplied by the value at point T, thus,

$$\mathbb{P}\left[T - \frac{T}{\ell^m} \leq \sum_{i=1}^{\ell} \Delta T_i < T - \frac{T}{\ell^{m+1}}\right] \leq C\ell^{-m} f(T) \leq C\ell^{-m} \frac{T^{\ell\kappa - 1} e^{-T/\eta}}{\eta^{\ell\kappa} \Gamma(\ell\kappa)} ,$$

where we have used the p.d.f. of a Gamma random variable to obtain the last inequality. Similar to the case $\ell = 1$ we can bound the expectation by

$$\begin{aligned}
&\mathbb{E}\left[\left(\beta \prod_{k=2}^{N_T} P_k\right)^2 \middle| N_T = \ell, \frac{T}{\ell^{m+1}} \leq \Delta T_{N_T+1} < \frac{T}{\ell^m}\right] \\
&\quad \leq \mathbb{E}\left[C^{N_T} \Delta T_{N_T+1}^{-1+\nu/2} \prod_{k=2}^{N_T} \Delta T_k^{-(2-\nu)n - 2\kappa} \middle| N_T = \ell, \frac{T}{\ell^{m+1}} \leq \Delta T_{N_T+1} < \frac{T}{\ell^m}\right] .
\end{aligned}$$

A simple requirement for the product to be bounded is $-(2-\nu)n - 2\kappa \geq 0$, by Assumption

13.2.5 $\kappa = 1/2$, hence $-n \geq 1/(2-\nu)$. As it turns out, taking $\nu = 1$ is useful to complete the proof, therefore we require $n \leq -1$, which holds by Assumption 13.2.5. This set of κ, ν and n also allow us to bound (13.2.12), hence we only considered (13.2.15).

The only term we have to consider in the expectation is $\Delta T_{N_T+1}^{-1+\nu/2}$, but by our conditioning this is bounded by $T\ell^{(1-\nu/2)(m+1)}$, hence for fixed $\ell \geq 2$ and letting $\nu = 1$ we obtain the following,

$$\mathbb{E}\left[\left(\beta\prod_{k=2}^{N_T} P_k\right)^2 \middle| N_T = \ell\right] \leq C^\ell \ell + \frac{1}{\mathbb{P}[N_T=\ell]}\sum_{m=1}^{\infty} C^\ell \ell^{(1/2)(m+1)}\ell^{-m}\frac{T^{\ell\kappa-1}e^{-T/\eta}}{\eta^{\ell\kappa}\Gamma(\ell\kappa)}\,.$$

One can easily see that the sum in m converges since $(1/2)(m+1)-m \leq 0$ for $m \geq 1$ and $\ell \geq 2$, the sum can be easily bounded by $\sum_{m=1}^{\infty} 2^{-(1/2)m+1/2} = C$ for any $\ell \geq 2$. One can compare this to the result in [DOW17, Proposition 4.1] where the authors obtain a bound of the form C^ℓ, hence our bound is not as strong but it is still good enough to ensure convergence.

Step 5: The sum over N_T converges. The final step of the proof is to show that the overall sum converges. We proceed by observing the following (see [DOW17, Proposition 4.1]),

$$\mathbb{P}[N_T = \ell] \leq \frac{C^{\ell\kappa}}{\ell\kappa\Gamma(\ell\kappa)}\,.$$

Using a generalisation of Stirling's formula one can approximate $\Gamma(z) \sim z^{z-1/2}e^{-z}\sqrt{2\pi}$. Hence we can bound

$$\begin{aligned}\mathbb{E}\left[\beta\prod_{k=2}^{N_T} P_k \mathbb{1}_{\{N_T\geq 1\}}\right] &\leq \sum_{\ell=1}^{\infty} C^\ell \ell \frac{C^{\ell\kappa}}{\ell\kappa\Gamma(\ell\kappa)} + \frac{\mathbb{P}[N_T=\ell]}{\mathbb{P}[N_T=\ell]}\sum_{m=1}^{\infty} C^\ell \ell^{(1/2)(m+1)}\ell^{-m}\frac{T^{\ell\kappa-1}e^{-T/\eta}}{\eta^{\ell\kappa}\Gamma(\ell\kappa)}\\ &\leq \sum_{\ell=1}^{\infty} C^\ell \frac{C^{\ell\kappa}}{\kappa\Gamma(\ell\kappa)}\,,\end{aligned}$$

and using Stirling's formula,

$$C^\ell\frac{C^{\ell\kappa}}{\kappa\Gamma(\ell\kappa)} \sim C^\ell\frac{C^{\ell\kappa}e^{\ell\kappa}}{\kappa(\ell\kappa)^{\ell\kappa-1/2}\sqrt{2\pi}} \leq \left(\frac{C^{1/\kappa}e^1}{\ell\kappa}\right)^{\ell\kappa-1/2} C^{1/(2\kappa)}e^{1/2}\,,$$

since $\kappa = 1/2$ this gives a sequence that converges under summation. □

Remark 13.2.9 (Optimal σ_0)**.** *One can see from the variance calculations that the $\frac{\sigma(\theta,T_0)^4}{\sigma(\theta,T_{N_T})^2}$ will leave a σ_0^2 term behind. Thus as one would expect the variance will be minimised by taking σ_0 smaller, however, to deal with terms involving nonlinearities in $\partial_x v$ one obtains terms of the form $\frac{1}{\sigma}$ thus an optimisation needs to be performed in order to set σ_0 at the correct level. Crucially however, the expected value (bias) is not effected by this choice.*

13.2.2 Estimator solves the PDE under enough integrability

At this point we have only proved that the estimator can be approximated via Monte Carlo. We now show that given some extra integrability conditions the estimator solves PDE (13.1.1). The final step is to show the said integrability conditions hold.

Theorem 13.2.10 is the analogous result to Theorem 3.5 in [HLOT$^+$16] (Theorem 12.1.7), however, the representation we derive below is more complex. The reason for the added complexity is the antithetic as well as the control variate on the final jump. Where as the control variate keeps the final Malliavin weight the same, the antithetic changes the weight, this then requires us to have extra terms that [HLOT$^+$16] does not have.

Theorem 13.2.10. *Let Assumptions 13.2.2, 13.2.3 and 13.2.5 hold. Define the following random* *variables,*

$$\begin{aligned}
\tilde{\psi}^{t,x} &:= \Big(\frac{\Delta g_{T_{N_T+1}}}{2\overline{F}(\Delta T_{N_T+1})}\frac{\Delta b_{N_T}\mathcal{W}^1_{N_T+1} - \frac{1}{2}\sigma(\theta_{N_T-1},T_{N_T})^2\mathcal{W}^2_{N_T+1}}{f(\Delta T_{N_T})} \\
&\quad + \frac{\Delta \hat{g}_{T_{N_T+1}}}{2\overline{F}(\Delta T_{N_T+1})}\frac{-\Delta b_{N_T}\mathcal{W}^1_{N_T+1} - \frac{1}{2}\sigma(\theta_{N_T-1},T_{N_T})^2\mathcal{W}^2_{N_T+1}}{f(\Delta T_{N_T})}\Big)\prod_{k=2}^{N_T}\frac{\Delta b_{k-1}\mathcal{W}^1_k - \frac{1}{2}\sigma(\theta_{k-2},T_{k-1})^2\mathcal{W}^2_k}{f(\Delta T_{k-1})}\,, \\
\psi^{t,x} &:= \mathbb{1}_{\{N_T=0\}}\frac{g(\bar{X}_{T_{N_T+1}})}{\overline{F}(\Delta T_{N_T+1})} + \mathbb{1}_{\{N_T\geq 1\}}\tilde{\psi}^{t,x}\,,
\end{aligned}$$

and

$$\Phi_1^{T_{N_T},\bar{X}_{T_{N_T}}} = \frac{\Delta g_{T_{N_T+1}} - \Delta\hat{g}_{T_{N_T+1}}}{2\overline{F}(\Delta T_{N_T+1})} \quad \textit{and} \quad \Phi_2^{T_{N_T},\bar{X}_{T_{N_T}}} = \frac{\Delta g_{T_{N_T+1}} + \Delta\hat{g}_{T_{N_T+1}}}{2\overline{F}(\Delta T_{N_T+1})}$$

where

$$\begin{aligned}
\Delta g_{T_{N_T+1}} &:= g(\bar{X}_{T_{N_T+1}}) - g(\bar{X}_{T_{N_T}} + b(T_{N_T})\Delta T_{N_T+1})\,, \\
\Delta \hat{g}_{T_{N_T+1}} &:= g(\hat{X}_{T_{N_T+1}}) - g(\bar{X}_{T_{N_T}} + b(T_{N_T})\Delta T_{N_T+1})\,,
\end{aligned}$$

the first and second order Malliavin weights are given by,

$$\mathcal{W}^1_{k+1} = \sigma(\theta_k, T_{k+1})^{-1}\frac{\Delta W_{T_{k+1}}}{\Delta T_{k+1}} \quad \textit{and} \quad \mathcal{W}^2_{k+1} = \sigma(\theta_k, T_{k+1})^{-2}\left(\frac{\Delta W^2_{T_{k+1}} - \Delta T_{k+1}}{\Delta T^2_{k+1}}\right). \tag{13.2.16}$$

The superscript in ψ, $\tilde{\psi}$, Φ_1 and Φ_2 denotes the initial condition for the SDE, $\bar{X}$. Further assume that,

$$\begin{aligned}
&\psi^{t,x},\ \tilde{\psi}^{t,x}\mathcal{W}^1_1,\ \tilde{\psi}^{t,x}\mathcal{W}^2_1,\ f(\Delta T_1)^{-1}\Delta b_1\tilde{\psi}^{T_1,\bar{X}_{T_1}}\mathcal{W}^1_2,\ f(\Delta T_1)^{-1}\sigma(\theta_0,T_1)^2\tilde{\psi}^{T_1,\bar{X}_{T_1}}\mathcal{W}^2_2, \\
&\Phi_1^{T_{N_T},\bar{X}_{T_{N_T}}}\mathcal{W}^1_{N_T+1},\ \Phi_2^{T_{N_T},\bar{X}_{T_{N_T}}}\mathcal{W}^2_{N_T+1}\,,
\end{aligned}$$

are uniformly integrable and that $\psi^{T_1,\bar{X}_{T_1}}$, $\Delta b_2\tilde{\psi}^{T_2,\bar{X}_{T_2}}\mathcal{W}^1_3$, $\sigma(\theta_1,T_2)^2\tilde{\psi}^{T_2,\bar{X}_{T_2}}\mathcal{W}^2_3$ are $\mathbb{P}$-a.s. uniformly integrable and $\tilde{\psi}^{T_1,\bar{X}_{T_1}}\mathcal{W}^1_2$ and $\tilde{\psi}^{T_1,\bar{X}_{T_1}}\mathcal{W}^2_2$ are $\mathbb{P}$-a.s. integrable.

Then, the function $\hat{v}(t,x) := \mathbb{E}[\psi^{t,x}|\mathcal{F}_t]$ solves the PDE (13.2.1).

Remark 13.2.11 ($\mathbb{P}$-a.s. (uniformly) integrable)**.** *Note that some of the processes stated in the theorem, for example $\psi^{T_1,\bar{X}_{T_1}}$ and $\tilde{\psi}^{T_1,\bar{X}_{T_1}}\mathcal{W}^2_2$ depend on random "initial conditions". Hence some of these processes are unbounded, but are finite up to a null set. For example, when we state $\tilde{\psi}^{T_1,\bar{X}_{T_1}}\mathcal{W}^2_2$ is $\mathbb{P}$-a.s. integrable, we mean that, $\mathbb{E}[|\tilde{\psi}^{T_1,\bar{X}_{T_1}}\mathcal{W}^2_2|\,|\mathcal{F}_{T_1}] < \infty$ $\mathbb{P}$-a.s. and similar for the uniform integrability condition.*

This theorem only shows that the estimator gives rise to the solution of the PDE under certain integrability assumptions. In order to finish our proof we need to show that such integrability conditions hold (Theorem 13.2.13). Although it is ψ that solves the PDE, our proof relies on various intermediary steps requiring additional integrability on $\psi\mathcal{W}$. Since one does not have this in general, we introduce the seemingly arbitrary $\tilde{\psi}$ and Φ which have the required integrability. Therefore, throughout the proof we show that one can view these additional processes as $\psi\mathcal{W}$ with a control variate and perform the various steps on $\tilde{\psi}$ and Φ.

Remark 13.2.12. *The Malliavin weights are given by* (13.1.8) *since our unbiased estimation puts us in the simple setting where the SDE has constant coefficients (see* [FLL+99]*).*

Proof. The main idea of this proof is to first show a stochastic representation for the PDE, then show that this representation and $\mathbb{E}[\psi^{t,x}|\mathcal{F}_t]$ are equivalent. Following Section 13.1.2, since a $C^{1,3}_b$ solution is assumed to exist, one can take constants b_0 and σ_0 and define the following PDE (equivalent to (13.2.1)),

$$\begin{cases}\partial_t v(t,x) + b_0 \partial_x v(t,x) + \frac{1}{2}\sigma_0^2 \partial_{xx} v(t,x) + (b(t) - b_0)\partial_x v(t,x) - \frac{1}{2}\sigma_0^2 \partial_{xx} v(t,x) = 0\,, \\ v(T,x) = g(x)\,. \end{cases}$$

Assume that these constants b_0 and σ_0 are adapted to the filtration $\mathcal{F}_t$ (as defined at the start of Section 13.2). Define $\tilde{X}$ as the solution to the SDE on $s \in [t,T]$

$$\mathrm{d}\tilde{X}_s = b_0 \mathrm{d}s + \sigma_0 \mathrm{d}W_s\,, \quad \tilde{X}_t = x\,.$$

again since $v \in C_b^{1,3}$, one obtains from the Feynman-Kac formula,

$$v(t,x) = \mathbb{E}\left[g(\tilde{X}_T) + \int_t^T (b(s) - b_0)\partial_x v(s, \tilde{X}_s) - \frac{1}{2}\sigma_0^2 \partial_{xx} v(s, \tilde{X}_s) \mathrm{d}s \,\Big|\mathcal{F}_t \right]\,.$$

It is important to note that we have not assigned values to the constants b_0 and σ_0 here, only that they are adapted to the initial filtration. Using standard branching arguments, we introduce a random variable independent of Brownian motion, corresponding to the life of the particle which allows us to rewrite the previous expression as[§],

$$v(t,x) = \mathbb{E}\left[\frac{g(\tilde{X}_T)}{\overline{F}(\Delta T_1)} \mathbb{1}_{\{T_1 = T\}} + \frac{\mathbb{1}_{\{T_1 < T\}}}{f(\Delta T_1)} \left\{ (b(T_1) - b_0)\partial_x v(T_1, \tilde{X}_{T_1}) - \frac{1}{2}\sigma_0^2 \partial_{xx} v(T_1, \tilde{X}_{T_1}) \right\} \Big|\mathcal{F}_t \right]\,. \tag{13.2.17}$$

As before, the representation does not depend on the value of the constants, therefore let us take $b_0 := b(t)$ and $\sigma_0 := \sigma_0$ (in the sense of (13.2.3)), thus $\tilde{X}$ is equivalent to $\bar{X}$.

This can be thought of as the forward representation, the goal now is to reach the same representation going backwards. Namely, starting from the estimator $\psi^{t,x}$, we want to remove the Malliavin weights and obtain the same relationship. We break the remainder of the proof into several steps.

Step 1: Continuity of the functions. We start by noting that between any two mesh points, the SDE is continuous w.r.t. its initial condition $(T_k, \bar{X}_{T_k})$, which is clear from the fact that it is just an SDE with constant coefficients. This along with the uniform integrability assumption of ψ implies that the function $\hat{v}$ is jointly continuous. This stems from the fact that we can define $\psi_n^{t,x}$ as $\psi^{t,x}$ but with the N_T replaced by $N_T \wedge n$, hence $\psi^{t,x} = \lim_{n\to\infty} \psi_n^{t,x}$. Then for each n we have a finite product of jointly continuous functions, which is therefore jointly continuous. Then uniform integrability allows us to take the limit as $n \to \infty$ inside to conclude that $(t,x) \to \mathbb{E}[\psi^{t,x}|\mathcal{F}_t]$ must also be a jointly continuous function.

The weights $\mathcal{W}^i$ are also continuous w.r.t. the initial condition. Thus by arguing in a similar way to above we have $\mathbb{E}[\tilde{\psi}^{t,x}\mathcal{W}_1^i|\mathcal{F}_t]$ and $\mathbb{E}[\Phi_i^{T_{N_T}, \bar{X}_{T_{N_T}}} \mathcal{W}_{N_T+1}^i|\mathcal{F}_t]$ are jointly continuous by the uniform integrability assumption.

Step 2: Rewriting the representation. By construction of ψ, there are two main cases, either the particle goes through a regime switch, which implies $\{N_T \geq 1\}$ or it "survives" until the end, $\{N_T = 0\}$. The key difference to the representation is the introduction of the variance reduction techniques when $\{N_T \geq 1\}$, this is also the distinction between ψ and $\tilde{\psi}$. Hence the

[§] There is a slight abuse of notation here, whereby $\mathbb{E}$ is now technically the expectation in the product space of the two random variables.

representation is,

$$\begin{aligned}\hat{v}(t,x) =&\mathbb{E}\Bigg[\mathbb{1}_{\{N_T=0\}}\frac{g(\bar{X}_{T_{N_T+1}})}{\overline{F}(\Delta T_{N_T+1})}\\
&\qquad+\mathbb{1}_{\{N_T\geq 1\}}\Bigg(\frac{\Delta g_{T_{N_T+1}}}{2\overline{F}(\Delta T_{N_T+1})}\frac{\Delta b_{N_T}\mathcal{W}^1_{N_T+1}-\frac{1}{2}\sigma(\theta_{N_T-1},T_{N_T})^2\mathcal{W}^2_{N_T+1}}{f(\Delta T_{N_T})}\\
&\qquad\qquad+\frac{\Delta\hat{g}_{T_{N_T+1}}}{2\overline{F}(\Delta T_{N_T+1})}\frac{-\Delta b_{N_T}\mathcal{W}^1_{N_T+1}-\frac{1}{2}\sigma(\theta_{N_T-1},T_{N_T})^2\mathcal{W}^2_{N_T+1}}{f(\Delta T_{N_T})}\Bigg)\\
&\qquad\times\prod_{k=2}^{N_T}\frac{\Delta b_{k-1}\mathcal{W}^1_k-\frac{1}{2}\sigma(\theta_{k-2},T_{k-1})^2\mathcal{W}^2_k}{f(\Delta T_{k-1})}\Bigg|X_t=x,\ \sigma(\theta_0,t)\Bigg],\end{aligned}$$

where we are using conditioning to state the initial condition of the SDE. In order to save space in the future we will stick to conditioning $\mathcal{F}_t$. Concentrating on the case $\{N_T\geq 1\}$, then the random variable ΔT_1 exists and satisfies $t<T_1<T$. Hence we can consider the filtration up to that point and by the tower property rewrite the $\{N_T\geq 1\}$ term in the expectation as,

$$\begin{aligned}\mathbb{E}\Bigg[&\mathbb{1}_{\{N_T\geq 1\}}\frac{1}{f(\Delta T_1)}\Bigg\{\\
&\Delta b_1\mathbb{E}\Bigg[\mathbb{1}_{\{N_T=1\}}\frac{\Delta g_{T_{N_T+1}}-\Delta\hat{g}_{T_{N_T+1}}}{2\overline{F}(\Delta T_{N_T+1})}\mathcal{W}^1_{N_T+1}\\
&\qquad+\mathbb{1}_{\{N_T>1\}}\Bigg(\frac{\Delta g_{T_{N_T+1}}}{2\overline{F}(\Delta T_{N_T+1})}\frac{\Delta b_{N_T}\mathcal{W}^1_{N_T+1}-\frac{1}{2}\sigma(\theta_{N_T-1},T_{N_T})^2\mathcal{W}^2_{N_T+1}}{f(\Delta T_{N_T})}\\
&\qquad\qquad+\frac{\Delta\hat{g}_{T_{N_T+1}}}{2\overline{F}(\Delta T_{N_T+1})}\frac{-\Delta b_{N_T}\mathcal{W}^1_{N_T+1}-\frac{1}{2}\sigma(\theta_{N_T-1},T_{N_T})^2\mathcal{W}^2_{N_T+1}}{f(\Delta T_{N_T})}\Bigg)\\
&\qquad\qquad\times\prod_{k=3}^{N_T}\frac{\Delta b_{k-1}\mathcal{W}^1_k-\frac{1}{2}\sigma(\theta_{k-2},T_{k-1})^2\mathcal{W}^2_k}{f(\Delta T_{k-1})}\mathcal{W}^1_2\Bigg|\mathcal{F}_{T_1}\Bigg]\\
-&\frac{1}{2}\sigma(\theta_0,T_1)^2\mathbb{E}\Bigg[\mathbb{1}_{\{N_T=1\}}\frac{\Delta g_{T_{N_T+1}}+\Delta\hat{g}_{T_{N_T+1}}}{2\overline{F}(\Delta T_{N_T+1})}\mathcal{W}^2_{N_T+1}\\
&\qquad+\mathbb{1}_{\{N_T>1\}}\Bigg(\frac{\Delta g_{T_{N_T+1}}}{2\overline{F}(\Delta T_{N_T+1})}\frac{\Delta b_{N_T}\mathcal{W}^1_{N_T+1}-\frac{1}{2}\sigma(\theta_{N_T-1},T_{N_T})^2\mathcal{W}^2_{N_T+1}}{f(\Delta T_{N_T})}\\
&\qquad\qquad+\frac{\Delta\hat{g}_{T_{N_T+1}}}{2\overline{F}(\Delta T_{N_T+1})}\frac{-\Delta b_{N_T}\mathcal{W}^1_{N_T+1}-\frac{1}{2}\sigma(\theta_{N_T-1},T_{N_T})^2\mathcal{W}^2_{N_T+1}}{f(\Delta T_{N_T})}\Bigg)\\
&\qquad\qquad\times\prod_{k=3}^{N_T}\frac{\Delta b_{k-1}\mathcal{W}^1_k-\frac{1}{2}\sigma(\theta_{k-2},T_{k-1})^2\mathcal{W}^2_k}{f(\Delta T_{k-1})}\mathcal{W}^2_2\Bigg|\mathcal{F}_{T_1}\Bigg]\quad\Bigg\}\Bigg|\mathcal{F}_t\Bigg],\end{aligned}\tag{13.2.18}$$

where we have used that Δb_1 and $\sigma(\theta_0,T_1)$ are bounded and our integrability assumptions on Φ and $\tilde{\psi}^{T_1,\bar{X}_{T_1}}$ to apply the tower property. We see here that the antithetic variable is causing extra difficultly since we need to treat the case $N_T=1$ separately.

Step 3: Existence and continuity of derivatives. In order to obtain the required expression we must also understand the derivatives of the function, hence we must show these derivatives exist and obtain a representation for them. One can identify the terms inside the conditional expectations as $\Phi_i^{T_{N_T},\bar{X}_{T_{N_T}}}\mathcal{W}^i_{N_T+1}$ and $\tilde{\psi}^{T_1,\bar{X}_{T_1}}\mathcal{W}^i_2$ for $i\in\{1,2\}$.

Let us denote by $\eta(T_1,\bar{X}_{T_1}):=\mathbb{E}[\psi^{t,x}|\mathcal{F}_{T_1}]$, notice that for the same reasons $\psi^{t,x}$ is a continuous function of x, $\eta(T_1,\bar{X}_{T_1})$ is continuous w.r.t. $\bar{X}_{T_1}$ (which is in turn continuous w.r.t. x). Let us now consider derivatives of this function w.r.t. x. However, one should note that this expectation is on the product space of random variables T_i and W. While the Malliavin automatic differentiation results only hold differentiating $\mathbb{E}_W[\cdot]$. Therefore we must swap the derivative with the expectation $\mathbb{E}_f$, which we have proved to be valid (actually shown a more general case) in Lemma 13.5.1 under the assumed integrability. Hence since we have a continuous function over a bounded interval, one can conclude via Lemma 13.5.1 and automatic differentiation,

$$\partial_x^i \hat{v}(t,x) = \partial_x^i \mathbb{E}\Big[\eta(T_1, \bar{X}_{T_1})\big|\mathcal{F}_t\Big] = \mathbb{E}\left[\eta(T_1, \bar{X}_{T_1})\mathcal{W}_1^i\big|\mathcal{F}_t\right] = \mathbb{E}\left[\psi^{t,x}\mathcal{W}_1^i\big|\mathcal{F}_t\right] .$$

Technically we have again used the Tower property to remove the final conditional expectation which requires integrability. We now show this is valid and due to the form of ψ we split into two terms,

$$\mathbb{E}\left[\psi^{t,x}\mathcal{W}_1^i\big|\mathcal{F}_t\right] = \mathbb{E}\left[\mathbb{1}_{\{N_T=0\}}\psi^{t,x}\mathcal{W}_1^i + \mathbb{1}_{\{N_T\geq 1\}}\psi^{t,x}\mathcal{W}_1^i\big|\mathcal{F}_t\right] .$$

One can automatically see that if $N_T \geq 1$ then $\psi = \tilde{\psi}$, for the case $N_T = 0$, we need to show equivalence between ψ and the corresponding Φ. Firstly let us show,

$$\mathbb{E}\left[\mathbb{1}_{\{N_T=0\}}\psi^{t,x}\mathcal{W}_1^1\big|\mathcal{F}_t\right] = \mathbb{E}\left[\mathbb{1}_{\{N_T=0\}}\Phi_1^{t,x}\mathcal{W}_1^1\big|\mathcal{F}_t\right] .$$

Expanding out Φ_1 we obtain,

$$\mathbb{E}\left[\mathbb{1}_{\{N_T=0\}}\Phi_1^{t,x}\mathcal{W}_1^1\big|\mathcal{F}_t\right] = \mathbb{E}\left[\mathbb{1}_{\{N_T=0\}}\frac{g(\bar{X}_{T_{N_T+1}}) - g(\hat{X}_{T_{N_T+1}})}{2\overline{F}(\Delta T_{N_T+1})}\mathcal{W}_1^1\Big|\mathcal{F}_t\right] .$$

Using that W and $-W$ have the same distribution and $\mathcal{W}^1$ is an odd function of the Brownian increment ΔW (see (13.2.16)) we obtain,

$$\mathbb{E}\left[\mathbb{1}_{\{N_T=0\}}\Phi_1^{t,x}\mathcal{W}_1^1\big|\mathcal{F}_t\right] = \mathbb{E}\left[2\mathbb{1}_{\{N_T=0\}}\frac{g(\bar{X}_{T_{N_T+1}})}{2\overline{F}(\Delta T_{N_T+1})}\mathcal{W}_1^1\Big|\mathcal{F}_t\right] ,$$

which shows the required result. Equivalently, we now show the equality

$$\mathbb{E}\left[\mathbb{1}_{\{N_T=0\}}\psi^{t,x}\mathcal{W}_1^2\big|\mathcal{F}_t\right] = \mathbb{E}\left[\mathbb{1}_{\{N_T=0\}}\Phi_2^{t,x}\mathcal{W}_1^2\big|\mathcal{F}_t\right] .$$

By a similar argument to above,

$$\begin{aligned}&\mathbb{E}\left[\mathbb{1}_{\{N_T=0\}}\Phi_2^{t,x}\mathcal{W}_1^2\big|\mathcal{F}_t\right]\\ &= \mathbb{E}\left[\mathbb{1}_{\{N_T=0\}}\frac{g(\bar{X}_{T_{N_T+1}}) + g(\hat{X}_{T_{N_T+1}}) - 2g(\bar{X}_{T_{N_T}} + b(T_{N_T})\Delta T_{N_T+1})}{2\overline{F}(\Delta T_{N_T+1})}\mathcal{W}_1^2\Big|\mathcal{F}_t\right] .\end{aligned}$$

By the fact that $g(\bar{X}_{T_{N_T}} + b(T_{N_T})\Delta T_{N_T+1})$ is $\overline{\mathcal{F}}_{N_T}$-adapted, and the weight has zero expectation we can remove this term from the expectation. Again, since W and $-W$ have the same distribution, and $\mathcal{W}^2$ is even we obtain,

$$\mathbb{E}\left[\mathbb{1}_{\{N_T=0\}}\Phi_2^{t,x}\mathcal{W}_1^2\big|\mathcal{F}_t\right] = \mathbb{E}\left[2\mathbb{1}_{\{N_T=0\}}\frac{g(\bar{X}_{T_{N_T+1}})}{2\overline{F}(\Delta T_{N_T+1})}\mathcal{W}_1^2\Big|\mathcal{F}_t\right] ,$$

again, this yields the required result. Thus the spatial derivatives of $\hat{v}$ satisfy,

$$\partial_x^i \hat{v}(t,x) = \mathbb{E}\left[\mathbb{1}_{\{N_T=0\}}\Phi_i^{t,x}\mathcal{W}_1^i + \mathbb{1}_{\{N_T\geq 1\}}\tilde{\psi}^{t,x}\mathcal{W}_1^i\big|\mathcal{F}_t\right] .$$

Uniform integrability of $\tilde{\psi}\mathcal{W}^i$ and $\Phi_i\mathcal{W}^i$ then implies $\partial_x^i \hat{v}(t,x)$ is a continuous function and one can use this integrability to also conclude $\partial_x^i \hat{v}(t,x) = \mathbb{E}\left[\psi^{t,x}\mathcal{W}_1^i\big|\mathcal{F}_t\right]$. Thus existence of the first and second spatial derivatives are assured.

Step 4: Representations match. Introducing the following notation, $N_T(s) := N_T - N_s$, i.e. the number of regime switches that occur between time s and T, with the obvious relation

$N_T(t) = N_T$.

To show that the two representations are the same, we need to consider the terms $\partial_x^i \hat{v}(T_1, \bar{X}_{T_1})$ for $t \le T_1 < T$. One has that,

$$\hat{v}(T_1, \bar{X}_{T_1}) = \mathbb{E}[\psi^{T_1, \bar{X}_{T_1}} | \mathcal{F}_{T_1}] \, .$$

To apply derivatives we again introduce the function $\eta(T_2, \bar{X}_{T_2}) = \mathbb{E}[\psi^{T_1, \bar{X}_{T_1}} | \mathcal{F}_{T_2}]$ and then Lemma 13.5.1 and Malliavin automatic differentiation implies,

$$\partial_x^i \hat{v}(T_1, \bar{X}_{T_1}) = \mathbb{E}[\psi^{T_1, \bar{X}_{T_1}} \mathcal{W}_2^i | \mathcal{F}_{T_1}] \quad \mathbb{P}\text{-a.s.}$$

Using the same arguments as before we can rewrite this as,

$$\partial_x^i \hat{v}(T_1, \bar{X}_{T_1}) = \mathbb{E}\left[\mathbb{1}_{\{N_T(T_1)=0\}} \Phi_i^{T_1, \bar{X}_{T_1}} \mathcal{W}_2^i + \mathbb{1}_{\{N_T(T_1)\ge 1\}} \tilde{\psi}^{T_1, \bar{X}_{T_1}} \mathcal{W}_2^i \middle| \mathcal{F}_{T_1} \right] \quad \mathbb{P}\text{-a.s.}$$

One then recognises the internal conditional expectations in (13.2.18) as the derivatives of $\hat{v}$ starting at time $(T_1, \bar{X}_{T_1})$. Thus, by integrability, (13.2.18) can be simply written as,

$$\mathbb{E}\left[\mathbb{1}_{\{N_T \ge 1\}} \frac{1}{f(\Delta T_1)} \left(\Delta b_1 \partial_x \hat{v}(T_1, \bar{X}_{T_1}) - \frac{1}{2} \sigma(\theta_0, T_1)^2 \partial_{xx} \hat{v}(T_1, \bar{X}_{T_1}) \right) \middle| \mathcal{F}_t \right] .$$

This leads us to the following nonlinear relation for $\hat{v}$,

$$\hat{v}(t,x) = \mathbb{E}\left[\frac{g(\bar{X}_{T_1})}{\overline{F}(\Delta T_1)} \mathbb{1}_{\{N_T = 0\}} + \mathbb{1}_{\{N_T \ge 1\}} \frac{\Delta b_1 \partial_x \hat{v}(T_1, \bar{X}_{T_1}) - \frac{1}{2}\sigma(\theta_0, T_1)^2 \partial_{xx} \hat{v}(T_1, \bar{X}_{T_1})}{f(\Delta T_1)} \middle| \mathcal{F}_t \right] .$$

Since this representation and (13.2.17) are equal we have $v(t,x) = \hat{v}(t,x)$ hence our representation solves the PDE. □

13.2.3 Verifying the integrability assumptions

Theorem 13.2.10 relied on various integrability assumptions and our final result is to show that these assumptions hold.

Theorem 13.2.13. *Let Assumptions 13.2.2, 13.2.3 and 13.2.5 hold. Then the integrability conditions in Theorem 13.2.10 hold.*

Proof. We start by showing the uniform integrability conditions, recall that for uniform integrability to hold it is sufficient to show the stochastic process is in L^p for $p > 1$ (see [Wil91, Chapter 13] for results on uniform integrability).

Firstly, by Proposition 13.2.8, one can conclude that $\psi^{t,x} \in L^2$, thus we have the required uniform integrability. Let us now consider $\tilde{\psi}^{t,x} \mathcal{W}_1^1$ and $\tilde{\psi}^{t,x} \mathcal{W}_1^2$. Due to both quantities having very similar forms we consider $\tilde{\psi}^{t,x} \mathcal{W}_1^i$ for $i \in \{1, 2\}$, hence we want to show,

$$\mathbb{E}[|\tilde{\psi}^{t,x} \mathcal{W}_1^i|^p | \mathcal{F}_t] < \infty, \quad \text{for some } p > 1.$$

We show this by borrowing many of the arguments in the proof of Proposition 13.2.8, hence we take $p = 2$. Using the representation for $\tilde{\psi}^{t,x}$ and taking common factors we obtain,

$$\begin{aligned}
\mathbb{E}[|\tilde{\psi}^{t,x} \mathcal{W}_1^i|^2 | \mathcal{F}_t] \le & \mathbb{E}\left[\left(\frac{\Delta g_{T_{N_T+1}} - \Delta \hat{g}_{T_{N_T+1}}}{2\overline{F}(\Delta T_{N_T+1})} \frac{\Delta b_{N_T} \mathcal{W}_{N_T+1}^1}{f(\Delta T_{N_T})} \right)^2 \prod_{k=2}^{N_T} P_k^2 \left(\mathcal{W}_1^i\right)^2 \middle| \mathcal{F}_t \right] \\
& + \mathbb{E}\left[\left(\frac{\Delta g_{T_{N_T+1}} + \Delta \hat{g}_{T_{N_T+1}}}{2\overline{F}(\Delta T_{N_T+1})} \frac{\frac{1}{2}\sigma(\theta_{N_T-1}, T_{N_T})^2 \mathcal{W}_{N_T+1}^2}{f(\Delta T_{N_T})} \right)^2 \prod_{k=2}^{N_T} P_k^2 \left(\mathcal{W}_1^i\right)^2 \middle| \mathcal{F}_t \right] .
\end{aligned}$$

We now use the same techniques from the proof of Proposition 13.2.8, firstly, we can condition on $N_T = \ell$ and multiply by the corresponding probability. Then by conditioning on $\overline{\mathcal{F}}_{N_T}$ (see proof of Proposition 13.2.8) we obtain the following,

$$\mathbb{E}\left[\left(\frac{\Delta g_{T_{N_T+1}} - \Delta \hat{g}_{T_{N_T+1}}}{2\overline{F}(\Delta T_{N_T+1})} \frac{\Delta b_{N_T} \mathcal{W}^1_{N_T+1}}{f(\Delta T_{N_T})}\right)^2 \Big| \, \overline{\mathcal{F}}_{N_T}, \; N_T = \ell\right] \le C \frac{\Delta b^2_{N_T}}{f(\Delta T_{N_T})^2} \, ,$$

and

$$\begin{aligned}
&\mathbb{E}\left[\left(\frac{\Delta g_{T_{N_T+1}} + \Delta \hat{g}_{T_{N_T+1}}}{2\overline{F}(\Delta T_{N_T+1})} \frac{\frac{1}{2}\sigma(\theta_{N_T-1}, T_{N_T})^2 \mathcal{W}^2_{N_T+1}}{f(\Delta T_{N_T})}\right)^2 \Big| \, \overline{\mathcal{F}}_{N_T}, \; N_T = \ell\right] \\
&\qquad \le \frac{C}{f(\Delta T_{N_T})^2} \frac{\sigma(\theta_{N_T-1}, T_{N_T})^4}{\sigma(\theta_{N_T}, T_{N_T+1})^2} \frac{\min\left[1, \sigma(\theta_{N_T}, T_{N_T+1})^2 \Delta T_{N_T+1}\right]}{\Delta T_{N_T+1}} \, .
\end{aligned}$$

We now use these bounds to bound $\tilde{\psi}\mathcal{W}$. Concentrating on the Δb_{N_T} term, we follow the finite variance proof and condition out $\Delta b^2_{N_T} P^2_{N_T}$, then use (13.2.11), namely,

$$\begin{aligned}
&\mathbb{E}\left[\left(\frac{\Delta g_{T_{N_T+1}} - \Delta \hat{g}_{T_{N_T+1}}}{2\overline{F}(\Delta T_{N_T+1})} \frac{\Delta b_{N_T} \mathcal{W}^1_{N_T+1}}{f(\Delta T_{N_T})}\right)^2 \prod_{k=2}^{N_T} P^2_k \left(\mathcal{W}^i_1\right)^2 \Big| \, \mathcal{F}_t, \; N_T = \ell\right] \\
&\qquad \le \mathbb{E}\left[\frac{C}{f(\Delta T_{N_T})^2} \frac{1}{f(\Delta T_{N_T-1})^2} \frac{\sigma(\theta_{N_T-2}, T_{N_T-1})^4}{\sigma(\theta_{N_T-1}, T_{N_T})^4} \prod_{k=2}^{N_T-1} P^2_k \left(\mathcal{W}^i_1\right)^2 \Big| \, \mathcal{F}_t, \; N_T = \ell\right] .
\end{aligned}$$

By continuing to follow the argument we can bound the above quantity by,

$$\mathbb{E}\left[\frac{C^{N_T}}{f(\Delta T_{N_T})^2} \frac{\sigma(\theta_0, T_1)^4}{\sigma(\theta_{N_T-1}, T_{N_T})^4} \prod_{k=2}^{N_T-1} \frac{1}{f(\Delta T_k)^2 \Delta T_k^2} \frac{1}{f(\Delta T_1)^2} \left(\mathcal{W}^i_1\right)^2 \Big| \, \mathcal{F}_t, \; N_T = \ell\right] . \qquad (13.2.19)$$

Since $\sigma_0 > 0$ is constant it is clear that,

$$\mathbb{E}[\left(\mathcal{W}^1_1\right)^2 | \overline{\mathcal{F}}_0] \le C \mathbb{E}[\left(\mathcal{W}^2_1\right)^2 | \overline{\mathcal{F}}_0] \le C \frac{1}{\Delta T_1^2} \, .$$

Hence we can bound (13.2.19),

$$\mathbb{E}\left[\frac{C^{N_T}}{f(\Delta T_{N_T})^2} \frac{\sigma(\theta_0, T_1)^4}{\sigma(\theta_{N_T-1}, T_{N_T})^4} \prod_{k=1}^{N_T-1} \frac{1}{f(\Delta T_k)^2 \Delta T_k^2} \Big| \, \mathcal{F}_t, \; N_T = \ell\right] \le \mathbb{E}\left[C^{N_T} \Big| \, \mathcal{F}_t, \; N_T = \ell\right] ,$$

where the inequality follows from our assumptions on f and σ.

Using this argument to deal with the extra Malliavin weight and the arguments in Proposition 13.2.8, we also obtain,

$$\begin{aligned}
&\mathbb{E}\left[\left(\frac{\Delta g_{T_{N_T+1}} + \Delta \hat{g}_{T_{N_T+1}}}{2\overline{F}(\Delta T_{N_T+1})} \frac{\frac{1}{2}\sigma(\theta_{N_T-1}, T_{N_T})^2 \mathcal{W}^2_{N_T+1}}{f(\Delta T_{N_T})}\right)^2 \prod_{k=2}^{N_T} P^2_k \left(\mathcal{W}^i_1\right)^2 \Big| \, \mathcal{F}_t\right] \\
&\qquad \le \mathbb{E}\left[C^{N_T} \frac{\sigma(\theta_{N_T}, T_{N_T+1})^\nu \Delta T^{\nu/2}_{N_T+1}}{\sigma(\theta_{N_T}, T_{N_T+1})^2} \Delta T^{-1}_{N_T+1} \prod_{k=1}^{N_T} \Delta T_k^{-2\kappa} \Bigg| N_T = \ell\right] , \quad \text{for } \nu \in [0, 2] \, .
\end{aligned}$$

The finiteness of these bounds follows directly from Proposition 13.2.8.

For the $f(\Delta T_1)^{-1} \Delta b_1 \tilde{\psi}^{T_1, \bar{X}_{T_1}} \mathcal{W}^1_2$ and $f(\Delta T_1)^{-1} \sigma(\theta_0, T_1)^2 \tilde{\psi}^{T_1, \bar{X}_{T_1}} \mathcal{W}^2_2$ terms, these follow automatically from Proposition 13.2.8.

For uniform integrability of $\Phi_1 \mathcal{W}^1$, take $p = 2$ as above. Then use Cauchy-Schwarz and the Lipschitz property of g, which yields $|\Delta g_{T_{N_T+1}} - \Delta \hat{g}_{T_{N_T+1}}| \le C\sigma(\theta_{N_T}, T_{N_T+1}) |\Delta W_{T_{N_T+1}}|$. One notes that the σ and ΔT terms cancel and hence finite.

Similarly, for $\Phi_2 \mathcal{W}^2$, again take $p = 2$ and use Cauchy-Schwarz along with (13.2.8). Again all terms cancel which implies this is also finite and hence uniformly integrable.

The final integrability results we require are all $\mathbb{P}$-a.s. results. We have $\psi^{T_1, \bar{X}_{T_1}}$, $\Delta b_2 \tilde{\psi}^{T_2, \bar{X}_{T_2}} \mathcal{W}^1_3$

and $\sigma(\theta_1,T_2)^2\tilde{\psi}^{T_2,\bar{X}_{T_2}}\mathcal{W}_3^2$ are $\mathbb{P}$-a.s. uniformly integrable, and $\tilde{\psi}^{T_1,\bar{X}_{T_1}}\mathcal{W}_2^1$ and $\tilde{\psi}^{T_1,\bar{X}_{T_1}}\mathcal{W}_2^2$ are $\mathbb{P}$-a.s. integrable. However, these follow from the arguments above along with the fact that $t<T_1<T_2$ $\mathbb{P}$-a.s. hence $\sigma(\theta_1,T_2)<\infty$ $\mathbb{P}$-a.s. Hence we have shown all the required integrability conditions to use Theorem 13.2.10. □

The proof of Theorem 13.2.7 follows in a straightforward way by combining these results.

Proof of Theorem 13.2.7. By letting Assumptions 13.2.2, 13.2.3 and 13.2.5 hold, then Theorems 13.2.10 and 13.2.13 imply that our estimator $\tilde{v}$ given in (13.2.5) solves the PDE (13.2.1).

Moreover, Proposition 13.2.8, implies that ψ is square integrable and hence of finite variance. □

13.3 Towards the general case and future work

The methodology presented in this work can be extended to accommodate PDEs of the form,

$$\begin{cases}\partial_t v(t,x)+b(t)\cdot Dv(t,x)+h(t,x)=0 \quad \text{for all } (t,x)\in[0,T)\times\mathbb{R}^d\,,\\ v(T,x)=g(x)\,,\end{cases} \tag{13.3.1}$$

where h is a nice function and we still have $v\in C_b^{1,3}$. As in the case of standard branching representations one introduces a further probability measure $\mathbb{P}_B$ on the space $\{0,1\}$, where 0 signifies the case the particles dies (this can be thought of as a v^0 term) at position $(T_k,\bar{X}_{T_k})$ and we evaluate h at this position.

13.3.1 Allowing b to have spatial dependence

Throughout this chapter we have made the assumption that the drift b does not depend on space. The main reason for this is to ensure finite variance. One can consider replacing Assumption 13.2.2, with $b:[0,T]\times\mathbb{R}\to\mathbb{R}$, satisfying $1/2$-Hölder in time, Lipschitz in space and uniformly bounded and most of the arguments presented still hold. The bound that changes and makes the arguments more difficult is (13.2.9), to see this let us observe how Δb is bounded under these new assumptions,

$$\begin{aligned}&\mathbb{E}\big[\Delta b_k^4|\overline{\mathcal{F}}_{k-1},N_T=\ell\big]\\&\qquad\le C\mathbb{E}\Big[(b(T_k,\bar{X}_{T_k})-b(T_k,\bar{X}_{T_{k-1}}))^4+(b(T_k,\bar{X}_{T_{k-1}})-b(T_{k-1},\bar{X}_{T_{k-1}}))^4|\overline{\mathcal{F}}_{k-1},N_T=\ell\Big]\,.\end{aligned}$$

For the second term we can use $1/2$-Hölder continuity in time of b, for the first term we can Lipschitz continuity in space to obtain,

$$\begin{aligned}\mathbb{E}[(b(T_k,\bar{X}_{T_k})-b(T_k,\bar{X}_{T_{k-1}}))^4|\overline{\mathcal{F}}_{k-1},N_T=\ell]\le&C\mathbb{E}[(\bar{X}_{T_k}-\bar{X}_{T_{k-1}})^4|\overline{\mathcal{F}}_{k-1},N_T=\ell]\\\le&C\mathbb{E}[(\Delta T_k+\sigma(\theta_{k-1},T_k)\Delta W_{T_k})^4|\overline{\mathcal{F}}_{k-1},N_T=\ell]\\\le&C\sigma(\theta_{k-1},T_k)^4\Delta T_k^2\,.\end{aligned}$$

Since σ is bounded from below we can conclude,

$$\mathbb{E}[\Delta b_k^4|\overline{\mathcal{F}}_{k-1},N_T=\ell]\le C\sigma(\theta_{k-1},T_k)^4\Delta T_k^2\,.$$

It is also straightforward to see the same bound applies if we take b Lipschitz in time. The bounds on M and V still have the form

$$\begin{aligned}\mathbb{E}[M_{k+1}^4|\overline{\mathcal{F}}_k,N_T=\ell]&\le C\frac{\Delta b_k^4}{\Delta T_{k+1}^2}\frac{1}{\sigma(\theta_k,T_{k+1})^4}\,,\\\mathbb{E}[V_{k+1}^4|\overline{\mathcal{F}}_k,N_T=\ell]&\le C\frac{\sigma(\theta_{k-1},T_k)^8}{\sigma(\theta_k,T_{k+1})^8}\frac{1}{\Delta T_{k+1}^4}\,,\end{aligned}$$

although one should note that we cannot use the Δb bound above in the M term since they are w.r.t. different conditional expectations. That being said though one can still observe where a problem arises by considering,

$$\begin{aligned}&\mathbb{E}\big[\mathbb{E}[P_{k+1}^4|\overline{\mathcal{F}}_k, N_T=\ell]\big|\overline{\mathcal{F}}_{k-1}, N_T=\ell\big]\\ &\leq C\mathbb{E}\Big[\frac{1}{f(\Delta T_k)^4}\Big(\frac{\Delta T_k^2}{\Delta T_{k+1}^2}\frac{\sigma(\theta_{k-1},T_k)^4}{\sigma(\theta_k,T_{k+1})^4}+\frac{\sigma(\theta_{k-1},T_k)^8}{\sigma(\theta_k,T_{k+1})^8}\frac{1}{\Delta T_{k+1}^4}\Big)\Big|\overline{\mathcal{F}}_{k-1}, N_T=\ell\Big].\end{aligned}$$

Whereas in the proof we can bound (13.2.9) by the term arising from the V (i.e. the V bound dominates the M bound), that is not the case here. To see this take $n=-1$ for the coefficient in the σ, we then obtain,

$$C\mathbb{E}\Big[\frac{1}{f(\Delta T_k)^4}\frac{\Delta T_k^6}{\Delta T_{k+1}^2}\Big(1+\frac{\Delta T_k^2}{\Delta T_{k+1}^2}\Big)\Big|\overline{\mathcal{F}}_{k-1}, N_T=\ell\Big].$$

Therefore the 1 (term arising from the M) is larger if $\Delta T_k < \Delta T_{k+1}$, hence we cannot dominate in the same way. As it turns out this a not a problem for obtaining (13.2.12), however, it does become an issue for obtaining (13.2.15). This appears because (13.2.15) relies on a cancelling argument, while this extra term changes the original bound from,

$$\mathbb{E}\Big[\frac{1}{f(\Delta T_k)^4}\frac{\Delta T_k^8}{\Delta T_{k+1}^4}\Big|\overline{\mathcal{F}}_{k-1}, N_T=\ell\Big] \quad \text{to} \quad \mathbb{E}\Big[\frac{1}{f(\Delta T_k)^4}\frac{\Delta T_k^8}{\Delta T_{k+1}^4\Delta T_k^2}\Big|\overline{\mathcal{F}}_{k-1}, N_T=\ell\Big].$$

This extra ΔT_k dependency makes the bound far weaker and consequently proving finite variance becomes more difficult. Of course the new bound we have obtained is not sharp, for example in the case $\Delta T_k \geq \Delta T_{k+1}$ we can return the original bound.

If we wish to argue the proof in a similar way one must either look to obtain a stronger bound on Δb (this is essentially why b in Assumption 13.2.2 worked), or one can find a way to make the V term dominate without increasing its size so much to break the remainder of the proof. For example, an interesting route to explore is to add an event probability distribution to the M and V term (similar to other branching diffusion algorithms) applying a judicious choice of probability distribution may give us the means to bound the M term by V again.

There are of course many different approaches one can take to solve this problem and as described, the remaining arguments in Theorems 13.2.10 and 13.2.13 follow with a more general b. But proving finite variance of this representation remains an open question.

13.3.2 Fully nonlinear first order case

Of course the true end goal of this work is to handle nonlinearities, for example, Burger's type vDv, which arise in many applications and for which numerical methods like characteristics cannot apply. Therefore future work will be on addressing explicit conditions under which this method provides solutions to transport PDEs of the form,

$$\begin{cases}\partial_t v(t,x)+b(t,x)\cdot Dv(t,x)=f(t,x,v,Dv)\,,\\ v(T,x)=g(x)\,,\end{cases}$$

where f is polynomial in v and Dv.

Handling such general first order PDEs will require additional arguments to what we have presented here. However, ideas from the case $b(t,x)$ along with the (purely numerical) technique presented in [War17] may yield the necessary tools to overcome such equations.

Remark 13.3.1 (Requirement for Smooth Solutions). *In theory this technique should be able to extend to the general, fully nonlinear case, one will still require a sufficiently smooth classical solution to the underlying PDE. The reason for this is due to the fact we assign a representation to $\partial_{xx}v$, thus we automatically require existence of this quantity.*

This implies that if we argue that the representation solves the PDE via viscosity solutions then we in fact show a classical solution. Of course this implies the method is not suitable for PDEs with *“shocks”.*

13.4 Examples

We show the potential of this method on two examples to compare this technique against the standard perturbation technique. The first example is a simple linear PDE which satisfies all of our assumptions and hence is only an example to show that our algorithm converges to the true, while the perturbation converges to a different value. The second is a nonlinear first order PDE, this is the more interesting case and we still observe our method giving reasonable results.

13.4.1 Simple First Order PDE

Let us consider the following linear PDE,

$$\begin{cases} \partial_t v(t,x) + \partial_x v(t,x) = 0 \quad \text{for all } (t,x) \in [0,1) \times \mathbb{R}\,, \\ v(1,x) = 10\cos(x-1-5)\,. \end{cases} \tag{13.4.1}$$

It is then clear to see that $v(t,x) = 10\cos(x-t-5)$ satisfies this PDE. Although such a PDE is easy to solve it serves as a good example to show the issue using a perturbation. We want to solve this PDE at the point $(0,10)$, where the true solution is ≈ 2.84. By considering the case where we perturb by $\sigma = 0.1$, and then estimate the expectation using varying amounts of Monte Carlo simulations, see Figure 13.1. To get a handle on the variance (error) we ran the simulation 50 times, plotted the average and the approximate 90% confidence interval. That is we view the largest and smallest value as a proxy for convergence of the algorithm. For the unbiased algorithm we also took, $n = -1$ and for the Gamma parameters $\kappa = 1/2$ and $\eta = 2$.

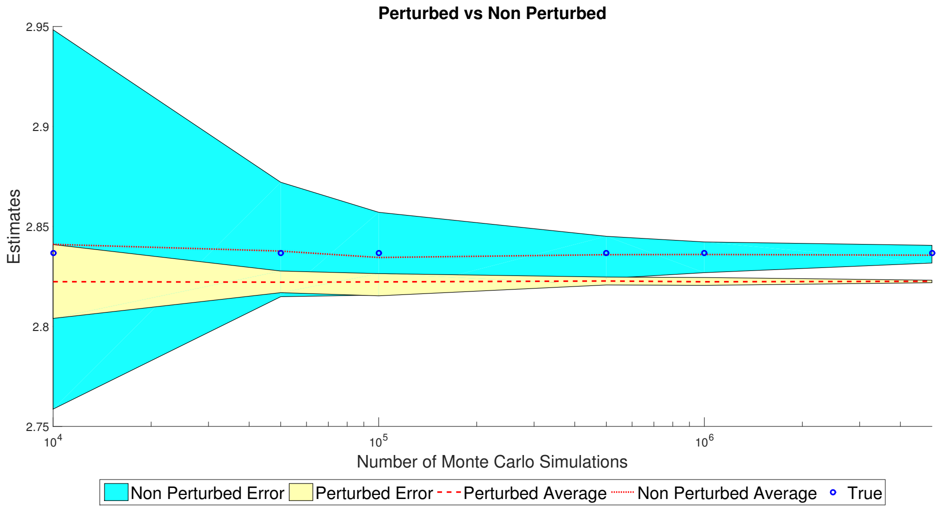


Figure 13.1: Shows the error and estimates of the solution as a function of the number of Monte Carlo simulations. The error corresponds to the approximate 90% confidence interval.

What is clear from Figure 13.1 is, as the number of Monte Carlo simulations increase, both algorithms are converging. However the perturbed case stays at a constant level away from the true value, which implies that the estimate is biased (as was expected). Therefore no amount of Monte Carlo simulations will yield the true solution. For the unbiased algorithm, although having a higher variance, we see that the average hovers around the true and moreover we observe convergence towards this point.

Hence the stochastic representation we derive indeed yields the true solution of the PDE, what is more fascinating and important about this result though is σ is not tending to zero, in fact we can bound it from below, this is the key step when it comes to more complex PDEs.

Moreover, this calculation was carried out using a basic Monte Carlo algorithm, one could look to more sophisticated techniques as appearing in [DOW17] where the authors apply particle methods for an improved convergence.

13.4.2 Nonlinear PDE

Let us now generalise to the nonlinear setting and consider the following PDE,

$$\begin{cases} \partial_t v(t,x) + \partial_x v(t,x) + \frac{1}{10}\big((\partial_x v(t,x))^2 + v(t,x)^2 - 1\big) = 0 & \text{for all } (t,x) \in [0,1) \times \mathbb{R}\,, \\ v(1,x) = \cos(1-x)\,. \end{cases} \tag{13.4.2}$$

We have taken this PDE since it is simple to observe that $v(t,x) = \cos(t-x)$ is the solution. It also is nice enough that one would expect our unbiased algorithm and the perturbation algorithm to work reasonably well. We want to solve this at the point $(0,1)$.

▷ *Convergence issue for the perturbation algorithm* One can note that, applying the perturbation technique implies that the resulting PDE is a second order semilinear PDE, and hence the corresponding branching algorithm is given in [HLOT$^+$16]. This creates a problem for the convergence of the algorithm, Assumption 3.10 and Theorem 3.12 of [HLOT$^+$16] give minimum bounds on the relative size of the drift to the diffusion, even for (13.4.2) which has a extremely nice solution, we observe that the algorithm fails to converge for $\sigma_0 = 0.5$ and has a large variance for σ_0 smaller than 1. Needless to say this is not a desirable property for the algorithm to have; perturbation can only work as a method if the perturbation is small and here we observe that there is a lower bound on the size of the perturbation and hence the bias of the estimator. Furthermore, as it turns out, there is no such problem with our unbiased algorithm and one can observe convergence for $\sigma_0 < 0.5$.

With the above in mind, in order to make the two algorithms comparable we set the perturbed algorithm as $\sigma_0 = 1$, but the remaining parameters are as above. Because the variance here is larger than the linear PDE we consider 100 realisations for each Monte Carlo level then take the approximate 80% confidence intervals and the average is then based on these 80 realisations. Furthermore, because we are dealing with nonlinear terms we have a more complex representation and need to establish a probability distribution for the type of event i.e. v^2, $(\partial_x v)^2$ etc. This is well understood in the case of the perturbation algorithm (see [HLOT$^+$16]), however, the variance of our unbiased algorithm seems to be highly dependent on how one chooses this probability distribution.

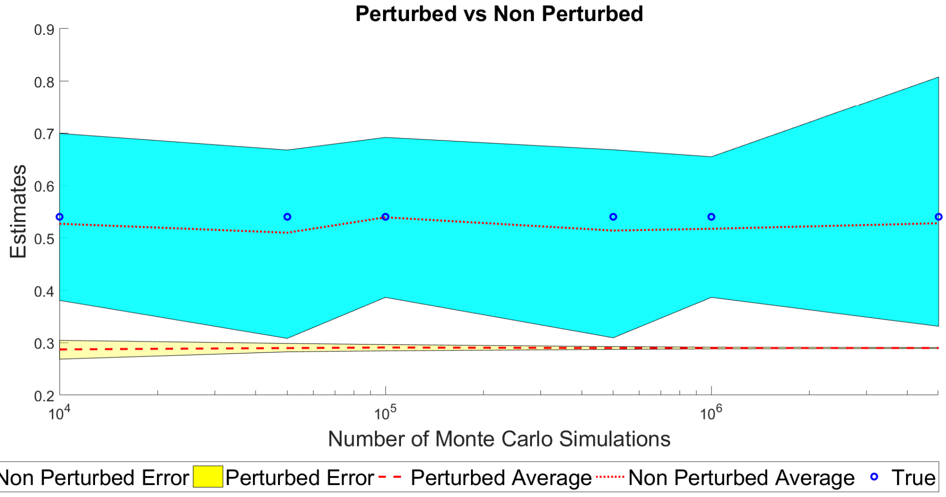


Figure 13.2: Shows the error and estimates of the solution as a function of the number of Monte Carlo simulations. The error corresponds to the approximate 80% confidence interval.

Figure 13.2 shows that yet again our unbiased algorithm provides a correction for the second order term. While the perturbation algorithm converges to a different value. However, it is clear that the variance in our algorithm is much higher. One of the reasons for this is because of the uncertainty in what events will be used for each realisation. Namely, for the linear PDE case, there was no probability distribution over events and this allowed us to bound the variance. In this more general case, more work would have to be done in order to bound the variance, and from our numerical example the choice of probability distribution has a role to play here.

13.5 Technical Result: Swapping Differentiation with Integration

When deriving the PDE we swapped the operators ∂_x with $\mathbb{E}_f$. This essentially requires taking a limit inside an integral, hence we show this is valid in this setting. A similar result was tackled in [HLTT17, Lemma A2], although our proof follows similar ideas to the one presented there, our version relaxes some of the conditions on the second derivative.

Lemma 13.5.1. *Let Assumptions 13.2.2, 13.2.3 and 13.2.5 hold. Moreover let $\psi^{T_1,\bar{X}_{T_1}}$, $\Delta b_2 \tilde{\psi}^{T_2,\bar{X}_{T_2}} \mathcal{W}_3^1$ and $\sigma(\theta_1, T_2)^2 \tilde{\psi}^{T_2,\bar{X}_{T_2}} \mathcal{W}_3^2$ be $\mathbb{P}$-a.s. uniformly integrable, let $\tilde{\psi}^{T_1,\bar{X}_{T_1}} \mathcal{W}_2^1$ and $\tilde{\psi}^{T_1,\bar{X}_{T_1}} \mathcal{W}_2^2$ be $\mathbb{P}$-a.s. integrable (as defined in Theorem 13.2.7), and define the function*

$$\hat{v}(T_1, \bar{X}_{T_1}) := \mathbb{E}_f[\mathbb{E}_W[\psi^{T_1,\bar{X}_{T_1}} | \mathcal{F}_{T_1}] | \mathcal{F}_{T_1}] \,.$$

Then for $i \in \{1, 2\}$,

$$\partial_x^i \hat{v}(T_1, \bar{X}_{T_1}) = \mathbb{E}_f[\partial_x^i \mathbb{E}_W[\psi^{T_1,\bar{X}_{T_1}} | \mathcal{F}_{T_1}] | \mathcal{F}_{T_1}] \quad \mathbb{P}\text{-a.s.}$$

Proof. Technically, the results below are for random variables and hence should be viewed in the a.s. sense, however, for ease of presentation we suppress writing a.s. at the end of each equation. Let us start by noting that,

$$\psi^{T_1,\bar{X}_{T_1}} = \mathbb{1}_{\{N_T(T_1)=0\}} \frac{g(\bar{X}_{T_{N_T+1}})}{\overline{F}(\Delta T_{N_T+1})} + \mathbb{1}_{\{N_T(T_1)\geq 1\}} \beta \prod_{k=3}^{N_T} P_k \,,$$

where $N_T(T_1) = N_T - N_{T_1}$. Observing that we can remove the time integral for the case $N_T(T_1) = 0$, that is,

$$\hat{v}(T_1, \bar{X}_{T_1}) = \mathbb{E}_f[\mathbb{E}_W[\mathbb{1}_{\{N_T(T_1)=0\}} \psi^{T_1,\bar{X}_{T_1}} + \mathbb{1}_{\{N_T(T_1)\geq 1\}} \psi^{T_1,\bar{X}_{T_1}} | \mathcal{F}_{T_1}] | \mathcal{F}_{T_1}] \,,$$

and by integrability we have

$$\begin{aligned} \mathbb{E}_f[\mathbb{E}_W[\mathbb{1}_{\{N_T(T_1)=0\}} \psi^{T_1,\bar{X}_{T_1}} | \mathcal{F}_{T_1}] | \mathcal{F}_{T_1}] &= \mathbb{E}_W[\mathbb{E}_f[\mathbb{1}_{\{N_T(T_1)=0\}} \psi^{T_1,\bar{X}_{T_1}} | \mathcal{F}_{T_1}] | \mathcal{F}_{T_1}] \\ &= \mathbb{E}_W[g(\bar{X}_{T_{N_T+1}}) | \mathcal{F}_{T_1}] \,. \end{aligned}$$

Hence we only need to consider the case $N_T(T_1) \geq 1$ hence $T_2 < T$. To make the proof easier we define the function φ for $T_1 < T_2 < T$ and $\bar{X}_{T_2} \in \mathbb{R}$ as follows,

$$\frac{1}{f(\Delta T_2)} \varphi^{T_1,\bar{X}_{T_1}}(T_2, \bar{X}_{T_2}) = \mathbb{E}[\mathbb{1}_{\{N_T(T_1)\geq 1\}} \psi^{T_1,\bar{X}_{T_1}} | \mathcal{F}_{T_2}] \,.$$

Following the argument as in Theorem 13.2.10 one can conclude from our uniform integrability assumption that for any $T_1 < T_2 < T$, $\varphi^{T_1,\bar{X}_{T_1}}(T_2, \bar{X}_{T_2})$ is $\mathbb{P}$-a.s. continuous in space i.e. w.r.t. $\bar{X}_{T_2}$. Further for any fixed $t < T_1 < T_2$, φ is bounded in space. To see this one can observe for

$T_2 < T$,

$$
\begin{aligned}
|\varphi^{T_1,\bar{X}_{T_1}}(T_2,\bar{X}_{T_2})| =& |f(\Delta T_2)\mathbb{E}[\mathbb{1}_{\{N_T(T_1)\geq 1\}}\psi^{T_1,\bar{X}_{T_1}}|\mathcal{F}_{T_2}]| \\
=& \Bigg| f(\Delta T_2)\mathbb{E}\Bigg[\Bigg(\frac{\Delta g_{T_{N_T+1}}}{2\overline{F}(\Delta T_{N_T+1})}\frac{\Delta b_{N_T}\mathcal{W}^1_{N_T+1} - \frac{1}{2}\sigma(\theta_{N_T-1},T_{N_T})^2\mathcal{W}^2_{N_T+1}}{f(\Delta T_{N_T})} \\
&+ \frac{\Delta \hat{g}_{T_{N_T+1}}}{2\overline{F}(\Delta T_{N_T+1})}\frac{-\Delta b_{N_T}\mathcal{W}^1_{N_T+1} - \frac{1}{2}\sigma(\theta_{N_T-1},T_{N_T})^2\mathcal{W}^2_{N_T+1}}{f(\Delta T_{N_T})}\Bigg) \\
&\times \prod_{k=3}^{N_T}\frac{\Delta b_{k-1}\mathcal{W}^1_k - \frac{1}{2}\sigma(\theta_{k-2},T_{k-1})^2\mathcal{W}^2_k}{f(\Delta T_{k-1})}\Bigg|\mathcal{F}_{T_2}\Bigg]\Bigg|.
\end{aligned}
$$

Removing $\mathcal{F}_{T_2}$-measurable terms and noticing that the remaining terms are integrable and $\Delta b_{k-1} < C$ independent of $\bar{X}_{T_2}$, we have $\varphi^{T_1,\bar{X}_{T_1}}(T_2,\cdot)$ is bounded in space, as required. Hence we can consider the following bounded Lipschitz approximation to φ,

$$
\varphi_n^{T_1,\bar{X}_{T_1}}(T_2,x) := \inf_{y\in\mathbb{R}}\Big\{\varphi^{T_1,\bar{X}_{T_1}}(T_2,y) + n|x-y|\Big\}\,.
$$

One can observe this approximation is both pointwise convergent and increasing in n. We therefore work with this approximation and take the limit to complete the proof.

Let us consider differentiating w.r.t. x, and in order to make all steps clear let us explicitly write each expectation. Using the tower property to write $\hat{v}$ in terms of φ then making the approximation we obtain,

$$
\begin{aligned}
&\partial_x \mathbb{E}_f\left[\mathbb{E}_W\left[\mathbb{1}_{\{N_T(T_1)\geq 1\}}\frac{1}{f(\Delta T_2)}\varphi_n^{T_1,\bar{X}_{T_1}}(T_2,\bar{X}_{T_2})\Big|\mathcal{F}_{T_1}\right]\Big|\mathcal{F}_{T_1}\right] \\
&= \lim_{\epsilon\to 0}\mathbb{E}_f\left[\frac{1}{\epsilon}\mathbb{E}_W\left[\mathbb{1}_{\{N_T(T_1)\geq 1\}}\frac{1}{f(\Delta T_2)}\big(\varphi_n^{T_1,\bar{X}_{T_1}+\epsilon}(T_2,\bar{X}^\epsilon_{T_2}) - \varphi_n^{T_1,\bar{X}_{T_1}}(T_2,\bar{X}_{T_2})\big)\Big|\mathcal{F}_{T_1}\right]\Big|\mathcal{F}_{T_1}\right]\,,
\end{aligned}
$$

where we are using the notation $\bar{X}^\epsilon_{T_2}$ to denote the SDE with initial condition perturbed by ϵ. Dominated convergence theorem implies we can take the limit inside the expectation if we show the "integrand" to be bounded. Using the Lipschitz assumption on φ_n, one has that,

$$
|\varphi_n^{T_1,\bar{X}_{T_1}+\epsilon}(T_2,\bar{X}^\epsilon_{T_2}) - \varphi_n^{T_1,\bar{X}_{T_1}}(T_2,\bar{X}_{T_2})| \leq C|\bar{X}^\epsilon_{T_2} - \bar{X}_{T_2}|\,.
$$

As stated in [HLTT17, Lemma A2], since $\bar{X}$ has constant coefficients the following bound holds,

$$
\mathbb{E}\left[\left|\frac{\bar{X}^\epsilon_{T_2} - \bar{X}_{T_2}}{\epsilon}\right|^2\Big|\mathcal{F}_{T_1}\right] \leq C\,, \tag{13.5.1}
$$

further, since $1/f(\Delta T_2) \leq C$ by dominated convergence theorem we can take the limit inside $\mathbb{E}_f$ to conclude,

$$
\begin{aligned}
&\partial_x \mathbb{E}_f\left[\mathbb{E}_W\left[\mathbb{1}_{\{N_T(T_1)\geq 1\}}\frac{\varphi_n^{T_1,\bar{X}_{T_1}}(T_2,\bar{X}_{T_2})}{f(\Delta T_2)}\Big|\mathcal{F}_{T_1}\right]\Big|\mathcal{F}_{T_1}\right] \\
&= \mathbb{E}_f\left[\partial_x\mathbb{E}_W\left[\mathbb{1}_{\{N_T(T_1)\geq 1\}}\frac{\varphi_n^{T_1,\bar{X}_{T_1}}(T_2,\bar{X}_{T_2})}{f(\Delta T_2)}\Big|\mathcal{F}_{T_1}\right]\Big|\mathcal{F}_{T_1}\right]\,.
\end{aligned}
$$

Completing the proof for the first derivative requires showing one can take the $\lim_{n\to\infty}$, however, we suppress this here and concentrate on the second derivative. One can check this holds by following the arguments presented in the case of the second derivative.

Again using the sequence of bounded Lipschitz functions we consider,

$$\partial_x^2 \mathbb{E}_f \left[\mathbb{E}_W \left[\mathbb{1}_{\{N_T(T_1)\geq 1\}} \frac{\varphi_n^{T_1,\bar{X}_{T_1}}(T_2,\bar{X}_{T_2})}{f(\Delta T_2)} \Big| \mathcal{F}_{T_1} \right] \Big| \mathcal{F}_{T_1} \right]$$
$$= \lim_{\epsilon\to 0} \mathbb{E}_f \left[\frac{1}{\epsilon} \mathbb{E}_W \left[\mathbb{1}_{\{N_T(T_1)\geq 1\}} \frac{\varphi_n^{T_1,\bar{X}_{T_1}+\epsilon}(T_2,\bar{X}^{\epsilon}_{T_2}) - \varphi_n^{T_1,\bar{X}_{T_1}}(T_2,\bar{X}_{T_2})}{f(\Delta T_2)} \mathcal{W}_2^1 \Big| \mathcal{F}_{T_1} \right] \Big| \mathcal{F}_{T_1} \right] ,$$

where we have used our first derivative result and the fact that φ_n is a bounded Lipschitz function to rewrite this derivative with a Malliavin weight. To bound this term one can apply Cauchy-Schwarz, use (13.5.1) and,

$$\mathbb{E}_W \left[\Big(\frac{\mathcal{W}_2^1}{f(\Delta T_2)} \Big)^2 \Big| \mathcal{F}_{T_1} \right] \leq C .$$

Hence we can again apply dominated convergence theorem to obtain,

$$\partial_x^2 \mathbb{E}_f \left[\mathbb{E}_W \left[\mathbb{1}_{\{N_T(T_1)\geq 1\}} \frac{\varphi_n^{T_1,\bar{X}_{T_1}}(T_2,\bar{X}_{T_2})}{f(\Delta T_2)} \Big| \mathcal{F}_{T_1} \right] \Big| \mathcal{F}_{T_1} \right]$$
$$= \mathbb{E}_f \left[\partial_x^2 \mathbb{E}_W \left[\mathbb{1}_{\{N_T(T_1)\geq 1\}} \frac{\varphi_n^{T_1,\bar{X}_{T_1}}(T_2,\bar{X}_{T_2})}{f(\Delta T_2)} \Big| \mathcal{F}_{T_1} \right] \Big| \mathcal{F}_{T_1} \right] .$$

To complete the proof we need to also take the $\lim_{n\to\infty}$, and have the expected values the same. Firstly recall that φ is an upperbound for φ_n, hence the result follows from the monotone convergence theorem (see [Wil91, Section 5.3]). Alternatively, one can use the upper bound and uniform integrability results in Theorem 13.2.13 to take the $\lim_{n\to\infty}$. □

13.6 Conclusions and Outlook

We have demonstrated a stochastic algorithm capable of dealing with first order PDEs, where originally such PDEs seemed beyond the reach of stochastic methods without approximation. This has potentially large implications for numerics of such PDEs since stochastic algorithms can easily be parallelised and scale favourable with dimension as argued in [BdRS17].

Future work focuses on the open questions left throughout this work. Namely assuming conditions that allow to show a solution to the PDE via its estimator, i.e. lifting Assumption 13.2.3. Secondly constructing a representation with finite variance that can deal with the full nonlinearities of the PDE, i.e. allowing for Burger's type nonlinearities on the RHS of PDE terms (13.2.1), or more generally fully nonlinear second order PDEs. Another interesting class related to this is degenerate second order semilinear PDEs which are also out of reach of the current theory. As mentioned there are multiple works being produced based on branching diffusions and numerical results seem to suggest branching algorithms are capable of handling fully nonlinear PDEs (see [War17]) although the theory is still under development. As branching diffusions (and regime switching) are relatively new areas the full potential of the techniques are still to be realised, consequently the somewhat restrictive assumptions in Chapter 12 can possibly be weakened to bring them more in-line with BSDEs. If this is the case, then branching diffusion can provide an extremely efficient (and potentially unbiased) high dimensional PDE solver.

With regard to finance the connection is straightforward, indeed many option pricing problems can be represented as a PDE and in particular if one has a basket option then these PDEs are high dimensional. Of course one does not always need the PDE representation, however, it is useful if one wishes to understand the surface of the solution. That is, stochastic representations are useful for solving at one single point and breaking the domain (as is done in PDD), however, if one wants to know the solution over multiple times or spatial positions then using the PDE representation is more efficient. Moreover, for much of the branching diffusion theory (in particular the unbiased results) it is easier to understand converting between the original expectation and the corresponding PDE. Related to this, one can also obtain the price of an Asian option as the solution to a degenerate second order PDE, this further emphasises why stochastic representations of PDEs are important in finance as well as many other fields.

Bibliography

[ABGK12] P. K Andersen, O. Borgan, R. D Gill, and N. Keiding, *Statistical models based on counting processes*, Springer Science & Business Media, 2012.

[ABLS05] J. A. Acebrón, M. P. Busico, P. Lanucara, and R. Spigler, *Domain decomposition solution of elliptic boundary-value problems via Monte Carlo and quasi-Monte Carlo methods*, SIAM J. Sci. Comput. **27** (2005), no. 2.

[AC17] A. Agarwal and J. Claisse, *Branching diffusion representation of quasi-linear elliptic PDEs and estimation using Monte Carlo method*, arXiv preprint arXiv:1704.00328 (2017).

[ADEH99] P. Artzner, F. Delbaen, J.-M. Eber, and D. Heath, *Coherent measures of risk*, Mathematical Finance. An International Journal of Mathematics, Statistics and Financial Economics **9** (1999), no. 3, 203–228. MR: 1850791 (2002d:91056).

[AHK91] P. K. Andersen, L. S. Hansen, and N. Keiding, *Non-and semi-parametric estimation of transition probabilities from censored observation of a non-homogeneous Markov process*, Scandinavian Journal of Statistics (1991), 153–167.

[AHL18] F. Ahmad, B. Hambly, and S. Ledger, *A stochastic partial differential equation model for the pricing of mortgage-backed securities*, Stochastic Processes and their Applications (2018).

[AK92] E. Altman and D. L. Kao, *The implications of corporate bond ratings drift*, Financial Analysts Journal **48** (1992), 64–75.

[AKH17] P. Andersson and A. Kohatsu-Higa, *Unbiased simulation of stochastic differential equations using parametrix expansions*, Bernoulli **23** (2017), no. 3, 2028–2057.

[ARRS09] J. A. Acebrón, Á. Rodríguez-Rozas, and R. Spigler, *Domain decomposition solution of nonlinear two-dimensional parabolic problems by random trees*, Journal of Computational Physics **228** (2009), no. 15, 5574–5591.

[ARRS10] J. A. Acebrón, Á. Rodríguez-Rozas, and R. Spigler, *A fully scalable algorithm suited for petascale computing and beyond*, Computer Science-Research and Development **25** (2010), no. 1-2, 115–121.

[BA16] F. Bernal and J. A. Acebrón, *A multigrid-like algorithm for probabilistic domain decomposition*, Computers & Mathematics with Applications **72** (2016), no. 7, 1790–1810.

[Bañ18] D. Baños, *The Bismut–Elworthy–Li formula for mean-field stochastic differential equations*, Annales de l'institut henri poincaré, probabilités et statistiques, 2018, pp. 220–233.

[BBC$^+$10] F. Bernardin, M. Bossy, C. Chauvin, J.-F. Jabir, and A. Rousseau, *Stochastic Lagrangian method for down-scaling problems in computational fluid dynamics*, ESAIM: Mathematical Modelling and Numerical Analysis **44** (2010), no. 5, 885–920.

[BCC11] F. Bolley, J. A. Cañizo, and J. A. Carrillo, *Stochastic mean-field limit: non-lipschitz forces and swarming*, Math Models Methods Appl Sci **21** (2011).

[BDF12] A. Budhiraja, P. Dupuis, and M. Fischer, *Large deviation properties of weakly interacting processes via weak convergence methods*, Ann. Probab. **40** (2012), no. 1, 74–102.

[BDK$^+$02] A. Bangia, F. X. Diebold, A. Kronimus, C. Schagen, and T. Schuermann, *Ratings migration and the business cycle, with application to credit portfolio stress testing*, Journal of Banking & Finance **26** (2002), no. 2, 445–474.

[BDL$^+$09] R. Buckdahn, B. Djehiche, J. Li, S. Peng, et al., *Mean-field backward stochastic differential equations: a limit approach*, The Annals of Probability **37** (2009), no. 4, 1524–1565.

[BdRS17] F. Bernal, G. dos Reis, and G. Smith, *Hybrid PDE solver for data-driven problems and modern branching*, European Journal of Applied Mathematics (2017), 1–24.

[BET04] B. Bouchard, I. Ekeland, and N. Touzi, *On the Malliavin approach to Monte Carlo approximation of conditional expectations*, Finance and Stochastics **8** (2004), no. 1, 45–71.

[BF17] A. Budhiraja and W.-T. Fan, *Uniform in time interacting particle approximations for nonlinear equations of Patlak-Keller-Segel type*, Electron. J. Probab. **22** (2017), Paper No. 8, 37. MR3613701

[BFFT12] J. Baladron, D. Fasoli, O. Faugeras, and J. Touboul, *Mean-field description and propagation of chaos in networks of Hodgkin-Huxley and FitzHugh-Nagumo neurons*, The Journal of Mathematical Neuroscience **2** (2012May), no. 1, 10.

[BFT15] M. Bossy, O. Faugeras, and D. Talay, *Clarification and complement to "mean-field description and propagation of chaos in networks of Hodgkin–Huxley and FitzHugh–Nagumo neurons"*, The Journal of Mathematical Neuroscience (JMN) **5** (2015), no. 1, 19.

[BG93] J. Besag and P. J. Green, *Spatial statistics and Bayesian computation*, Journal of the Royal Statistical Society. Series B (Methodological) (1993), 25–37.

[BJGR17] E. Bernton, P. E. Jacob, M. Gerber, and C. P. Robert, *Inference in generative models using the Wasserstein distance*, arXiv:1701.05146 (2017).

[BLP09] R. Buckdahn, J. Li, and S. Peng, *Mean-field backward stochastic differential equations and related partial differential equations*, Stochastic Processes and their Applications **119** (2009), no. 10, 3133–3154.

[BLP$^+$17] R. Buckdahn, J. Li, S. Peng, C. Rainer, et al., *Mean-field stochastic differential equations and associated pdes*, The Annals of Probability **45** (2017), no. 2, 824–878.

[BLR17] J. Bielagk, A. Lionnet, and G. D. Reis, *Equilibrium pricing under relative performance concerns*, SIAM Journal on Financial Mathematics **8** (2017), no. 1, 435–482.

[BMM15] E. Bacry, I. Mastromatteo, and J.-F. Muzy, *Hawkes processes in finance*, 2015.

[BMNS02] M. Bladt, B. Meini, M. F. Neuts, and B. Sericola, *Distributions of reward functions on continuous-time Markov chains*, Matrix-analytic methods (2002), 39–62.

[Bos04] M. Bossy, *Optimal rate of convergence of a stochastic particle method to solutions of 1d viscous scalar conservation laws*, Mathematics of computation **73** (2004), no. 246, 777–812.

[BS05] M. Bladt and M. Sorensen, *Statistical inference for discretely observed Markov jump processes*, Journal of the Royal Statistical Society: Series B (Statistical Methodology) **67** (2005), no. 3, 395–410.

[BS09] M. Bladt and M. Sorensen, *Efficient estimation of transition rates between credit ratings from observations at discrete time points*, Quantitative Finance **9** (2009), no. 2, 147–160.

[BT04] B. Bouchard and N. Touzi, *Discrete-time approximation and Monte-Carlo simulation of Backward Stochastic Differential Equations*, Stochastic Processes and their Applications **111** (2004), no. 2, 175–206.

[BT97] M. Bossy and D. Talay, *A stochastic particle method for the Mckean-Vlasov and the Burgers equation*, Mathematics of Computation of the American Mathematical Society **66** (1997), no. 217, 157–192.

[BTW17] B. Bouchard, X. Tan, and X. Warin, *Numerical approximation of general Lipschitz BSDEs with branching processes*, arXiv:1710.10933 (2017).

[Can04] R. Cantor, *An introduction to recent research on credit ratings*, Journal of Banking & Finance **28** (2004), no. 11, 2565–2573.

[Car08] R. Carmona, *Indifference pricing: theory and applications*, Princeton University Press, 2008.

[Car10] P. Cardaliaguet, *Notes on mean field games*, Technical report, 2010.

[Car16] R. Carmona, *Lectures on BSDEs, Stochastic Control, and Stochastic Differential Games with Financial Applications*, Vol. 1, SIAM, 2016.

[CCD14] J.-F. Chassagneux, D. Crisan, and F. Delarue, *A probabilistic approach to classical solutions of the master equation for large population equilibria*, arXiv preprint arXiv:1411.3009 (2014).

[CCP12] M. Choudhry, J. Cummins, and I. Plenderleith, *The principles of banking*, Vol. 5, Wiley Online Library, 2012.

[CD17a] R. Carmona and F. Delarue, *Probabilistic theory of mean field games with applications I*, 1st ed., Probability Theory and Stochastic Modelling, vol. 84, Springer International Publishing, 2017.

[CD17b] R. Carmona and F. Delarue, *Probabilistic theory of mean field games with applications II*, 1st ed., Probability Theory and Stochastic Modelling, vol. 84, Springer International Publishing, 2017.

[CDL13] R. Carmona, F. Delarue, and A. Lachapelle, *Control of McKean-Vlasov dynamics versus mean field games*, Mathematics and Financial Economics **7** (2013), no. 2, 131–166.

[CDS10] R. Cont, R. Deguest, and G. Scandolo, *Robustness and sensitivity analysis of risk measurement procedures*, Quantitative Finance **10** (2010), no. 6, 593–606.

[CD$^+$15] R. Carmona, F. Delarue, et al., *Forward-backward stochastic differential equations and controlled McKean-Vlasov dynamics*, The Annals of Probability **43** (2015), no. 5, 2647–2700.

[CH01] M. Carey and M. Hrycay, *Parameterizing credit risk models with rating data*, Journal of Banking & Finance **25** (2001), no. 1, 197 –270.

[CH08] G. Claeskens and N. L. Hjort, *Model selection and model averaging*, Cambridge Books (2008).

[CH12] B. Chen and Y. Hong, *Testing for the Markov property in time series*, Econometric Theory **28** (2012), no. 1, 130–178.

[CHL04] J. H. E. Christensen, E. Hansen, and D. Lando, *Confidence sets for continuous-time rating transition probabilities*, Journal of Banking & Finance **28** (2004), no. 11, 2575–2602.

[CIL92] M. G Crandall, H. Ishii, and P.-L. Lions, *User's guide to viscosity solutions of second order partial differential equations*, Bulletin of the American mathematical society **27** (1992), no. 1, 1–67.

[CJM16] J.-F. Chassagneux, A. Jacquier, and I. Mihaylov, *An explicit Euler scheme with strong rate of convergence for financial SDEs with non-Lipschitz coefficients*, SIAM Journal on Financial Mathematics **7** (2016), no. 1, 993–1021.

[CL83] M. G Crandall and P.-L. Lions, *Viscosity solutions of Hamilton-Jacobi equations*, Transactions of the American mathematical society **277** (1983), no. 1, 1–42.

[CM10] D. Crisan and K. Manolarakis, *Probabilistic methods for semilinear partial differential equations. Applications to finance*, Mathematical Modelling and Numerical Analysis **44** (2010), no. 5, 1107.

[CM17a] D. Crisan and E. McMurray, *Cubature on Wiener space for McKean–Vlasov SDEs with smooth scalar interaction*, 2017. arXiv:1703.04177.

[CM17b] D. Crisan and E. McMurray, *Smoothing properties of McKean–Vlasov SDEs*, Probability Theory and Related Fields (2017), 1–52.

[CMR05] O. Cappé, E. Moulines, and T. Rydén, *Inference in hidden Markov models*, Springer Verlag, New York, 2005.

[CO84] D. R. Cox and D. Oakes, *Analysis of survival data*, Routledge, 1984.

[Cou08] F. Couderc, *Credit Risk and Ratings: Understanding Dynamics and Relationships with Macroeconomics*, Ph.D. Thesis, 2008.

[CS02] B. A Craig and P. P Sendi, *Estimation of the transition matrix of a discrete-time Markov chain*, Health economics **11** (2002), no. 1, 33–42.

[CSTV07] P. Cheridito, H. M. Soner, N. Touzi, and N. Victoir, *Second-order Backward Stochastic Differential Equations and fully nonlinear parabolic PDEs*, Comm. Pure Appl. Math. **60** (2007), no. 7, 1081–1110.

[CT17] C. Cuchiero and J. Teichmann, *Stochastic representations of ordinary differential equations via affine processes*, 2017. Working paper.

[Cul66] W. J Culver, *On the existence and uniqueness of the real logarithm of a matrix*, Proceedings of the American Mathematical Society **17** (1966), no. 5, 1146–1151.

[Cut73] J. R. Cuthbert, *The logarithm function for finite-state Markov semi-groups*, Journal of the London Mathematical Society **2** (1973), no. 3, 524–532.

[CX10] D. Crisan and J. Xiong, *Approximate McKean–Vlasov representations for a class of SPDEs*, Stochastics An International Journal of Probability and Stochastics Processes **82** (2010), no. 1, 53–68.

[DE11] P. Dupuis and R. S. Ellis, *A weak convergence approach to the theory of large deviations*, Vol. 902, John Wiley & Sons, 2011.

[DFG$^+$16] W. Dreyer, P. K. Friz, P. Gajewski, C. Guhlke, and M. Maurelli, *Stochastic model for LFP-electrodes*, WIAS preprint no. 2329, 2016.

[DG87] D. A. Dawson and J. Gärtner, *Large deviations from the McKean-Vlasov limit for weakly interacting diffusions*, Stochastics **20** (1987), no. 4, 247–308.

[DGG$^+$11] W. Dreyer, M. Gaberšček, C. Guhlke, R. Huth, and J. Jamnik, *Phase transition in a rechargeable lithium battery*, European Journal of Applied Mathematics **22** (2011), no. 3, 267–290.

[DIR$^+$15] F. Delarue, J. Inglis, S. Rubenthaler, E. Tanré, et al., *Global solvability of a networked integrate-and-fire model of mckean–vlasov type*, The Annals of Applied Probability **25** (2015), no. 4, 2096–2133.

[DJM16] G. D'Amico, J. Janssen, and R. Manca, *Downward migration credit risk problem: a non-homogeneous backward semi-Markov reliability approach*, Journal of the Operational Research Society **67** (2016), no. 3, 393–401.

[DM13] P. Del Moral, *Mean field simulation for monte carlo integration*, CRC Press, 2013.

[DMG98] P. Del Moral and A. Guionnet, *Large deviations for interacting particle systems: applications to non-linear filtering*, Stochastic Process. Appl. **78** (1998), no. 1, 69–95. MR1653296

[DNØP09] G. Di Nunno, B. K. Øksendal, and F. Proske, *Malliavin calculus for Lévy processes with applications to finance*, Vol. 2, Springer, 2009.

[DOW17] M. Doumbia, N. Oudjane, and X. Warin, *Unbiased monte carlo estimate of stochastic differential equations expectations*, ESAIM: Probability and Statistics **21** (2017), 56–87.

[dRS17] G dos Reis and G Smith, *Robust and consistent estimation of generators in credit risk*, Quantitative Finance (2017), 1–19.

[dRST17] G. dos Reis, W. Salkeld, and J. Tugaut, *Freidlin-Wentzell LDPs in path space for McKean-Vlasov equations and the functional iterated logarithm law*, 2017. arXiv:1708.04961.

[Dud18] R. M Dudley, *Real Analysis and Probability: 0*, Chapman and Hall/CRC, 2018.

[DVJ03] D. J. Daley and D. Vere-Jones, *An introduction to the theory of point processes: Volume I: elementary theory and methods*, 2003.

[DVJ07] D. J. Daley and D. Vere-Jones, *An introduction to the theory of point processes: Volume II: general theory and structure*, Springer Science & Business Media, 2007.

[DW04] P. Dupuis and H. Wang, *Importance sampling, large deviations, and differential games*, Stochastics: An International Journal of Probability and Stochastic Processes **76** (2004), no. 6, 481–508.

[DY07] D. Dehay and J.-F. Yao, *On likelihood estimation for discretely observed Markov jump processes*, Australian & New Zealand Journal of Statistics **49** (2007), no. 1, 93–107.

[DZ10] A. Dembo and O. Zeitouni, *Large deviations techniques and applications, volume 38 of stochastic modelling and applied probability*, Springer-Verlag, Berlin, 2010.

[DZ13] A. Dassios and H. Zhao, *Exact simulation of Hawkes process with exponentially decaying intensity*, Electronic Communications in Probability **18** (2013), no. 62, 1–13.

[EKPQ97] N. El Karoui, S. Peng, and M. C. Quenez, *Backward stochastic differential equations in finance*, Math. Finance **7** (1997), no. 1, 1–71. MR1434407

[EKT+14] I. Ekren, C. Keller, N. Touzi, J. Zhang, et al., *On viscosity solutions of path dependent PDEs*, The Annals of Probability **42** (2014), no. 1, 204–236.

[ET99] I. Ekeland and R. Temam, *Convex analysis and variational problems*, SIAM, 1999.

[Eva98] L. C. Evans, *Partial Differential Equations*, American Mathematical Society, Providence, R.I, 1998.

[FBSN16] J. J Forster, M. Buzzacchi, A. Sudjianto, and R. Nagao, *Modelling credit grade migration in large portfolios using cumulative t-link transition models*, European Journal of Operational Research **254** (2016), no. 3, 977–984.

[Fer14] J.-D. Fermanian, *The limits of granularity adjustments*, Journal of Banking & Finance **45** (2014), 9–25.

[Fis14] M. Fischer, *On the form of the large deviation rate function for the empirical measures of weakly interacting systems*, Bernoulli **20** (2014), no. 4, 1765–1801.

[FLL+99] E. Fournié, J.-M. Lasry, J. Lebuchoux, P.-L. Lions, and N. Touzi, *Applications of malliavin calculus to monte carlo methods in finance*, Finance and Stochastics **3** (1999), no. 4, 391–412.

[FR75] W. H. Fleming and R. W. Rishel, *Deterministic and stochastic optimal control*, Springer-Verlag, Berlin-New York, 1975. Applications of Mathematics, No. 1.

[Fri12] A. Friedman, *Stochastic differential equations and applications*, Courier Corporation, 2012.

[FS08] H. Frydman and T. Schuermann, *Credit rating dynamics and Markov mixture models*, Journal of Banking & Finance **32** (2008), no. 6, 1062–1075.

[GC93] A. E. Gelfand and B. P. Carlin, *Maximum-likelihood estimation for constrained-or missing-data models*, Canadian Journal of Statistics **21** (1993), no. 3, 303–311.

[GDF01] N. Gordon, A. Doucet, and J. Freitas, *Sequential Monte Carlo methods in practice*, Springer-Verlag, 2001.

[GFB97] G. M. Gupton, C. C. Finger, and M. Bhatia, *Creditmetrics: technical document*, JP Morgan & Co., 1997.

[GHL13] J. Guyon and P. Henry-Labordère, *Nonlinear option pricing*, CRC Press, 2013.

[GHS99] P. Glasserman, P. Heidelberger, and P. Shahabuddin, *Asymptotically optimal importance sampling and stratification for pricing path-dependent options*, Mathematical finance **9** (1999), no. 2, 117–152.

[Gla13] P. Glasserman, *Monte Carlo methods in financial engineering*, Vol. 53, Springer Science & Business Media, 2013.

[GLL11] O. Guéant, J.-M. Lasry, and P.-L. Lions, *Mean field games and applications*, Paris-princeton lectures on mathematical finance 2010, 2011, pp. 205–266.

[GLS00] C. Gouriéroux, J. P. Laurent, and O. Scaillet, *Sensitivity analysis of values at risk*, Journal of empirical finance **7** (2000), no. 3, 225–245.

[GP15] E. Gobet and S. Pagliarani, *Analytical approximations of BSDEs with nonsmooth driver*, SIAM Journal on Financial Mathematics **6** (2015), no. 1, 919–958.

[GP18] E. Gobet and S. Pagliarani, *Analytical approximations of non-linear SDEs of McKean-Vlasov type*, Journal of Mathematical Analysis and Applications (2018).

[GR08] P. Guasoni and S. Robertson, *Optimal importance sampling with explicit formulas in continuous time*, Finance and Stochastics **12** (2008), no. 1, 1–19.

[GRS96] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, *Introducing Markov chain Monte Carlo*, Markov chain Monte Carlo in practice **1** (1996), 19.

[GT16] E. Gobet and P. Turkedjiev, *Linear regression MDP scheme for discrete backward stochastic differential equations under general conditions*, Mathematics of Computation **85** (2016), no. 299, 1359–1391.

[GW97] P. Glasserman and Y. Wang, *Counterexamples in importance sampling for large deviations probabilities*, The Annals of Applied Probability **7** (1997), no. 3, 731–746.

[Has70] W K. Hastings, *Monte Carlo sampling methods using Markov chains and their applications* (1970).

[HAT18] A.-L. Haji-Ali and R. Tempone, *Multilevel and Multi-index Monte Carlo methods for the McKean–Vlasov equation*, Statistics and Computing **28** (2018), no. 4, 923–935.

[HJ11] A. Hobolth and J. L. Jensen, *Summary statistics for endpoint-conditioned continuous-time markov chains*, Journal of applied probability **48** (2011), no. 4, 911–924.

[HJK11] M. Hutzenthaler, A. Jentzen, and P. E Kloeden, *Strong and weak divergence in finite time of Euler's method for stochastic differential equations with non-globally Lipschitz continuous coefficients*, Proceedings of the royal society of london a: Mathematical, physical and engineering sciences, 2011, pp. 1563–1576.

[HJK12] M. Hutzenthaler, A. Jentzen, and P. E Kloeden, *Strong convergence of an explicit numerical method for SDEs with nonglobally Lipschitz continuous coefficients*, The Annals of Applied Probability **22** (2012), no. 4, 1611–1641.

[HJK16] M. Hutzenthaler, A. Jentzen, and T. Kruse, *On full history recursive multilevel Picard approximations and numerical approximations of high-dimensional nonlinear parabolic partial differential equations*, arXiv preprint arXiv:1607.03295 (2016).

[HL12] P. Henry-Labordere, *Counterparty risk valuation: A marked branching diffusion approach*, SSRN 1995503 (2012).

[HLOT$^+$16] P. Henry-Labordere, N. Oudjane, X. Tan, N. Touzi, and X. Warin, *Branching diffusion representation of semilinear PDEs and Monte Carlo approximation*, arXiv:1603.01727 (2016).

[HLT18] P. Henry-Labordere and N. Touzi, *Branching diffusion representation for nonlinear Cauchy problems and Monte Carlo approximation*, arXiv preprint arXiv:1801.08794 (2018).

[HLTT14] P. Henry-Labordere, X. Tan, and N. Touzi, *A numerical algorithm for a class of BSDEs via the branching process*, Stochastic Processes and their Applications **124** (2014), no. 2, 1112–1140.

[HLTT17] P. Henry-Labordere, X. Tan, and N. Touzi, *Unbiased simulation of stochastic differential equations*, The Annals of Applied Probability **27** (2017), no. 6, 3305–3341.

[HMS02] D. J Higham, X. Mao, and A. M Stuart, *Strong convergence of Euler-type methods for nonlinear stochastic differential equations*, SIAM Journal on Numerical Analysis **40** (2002), no. 3, 1041–1063.

[HN16] H. Hult and P. Nyquist, *Large deviations for weighted empirical measures arising in importance sampling*, Stochastic Processes and their Applications **126** (2016), no. 1, 138–170.

[IJJK14] H. Inanoglu, M. Jacobs Jr, and A. K Karagozoglu, *Bank capital and new regulatory requirements for risks in trading portfolios*, The Journal of Fixed Income **23** (2014), no. 4, 71–88.

[Ina06] Y. Inamura, *Estimating continuous time transition matrices from discretely observed data*, Citeseer, 2006. (No. 06-E-7). Bank of Japan.

[IRW01] R. B. Israel, J. S. Rosenthal, and J. Z. Wei, *Finding generators for Markov chains via empirical transition matrices, with applications to credit ratings*, Mathematical finance **11** (2001), no. 2, 245–265.

[JLT97] R. A. Jarrow, D. Lando, and S. M. Turnbull, *A Markov model for the term structure of credit risk spreads*, Review of Financial studies **10** (1997), no. 2, 481–523.

[Kat75] T. Kato, *The Cauchy problem for quasi-linear symmetric hyperbolic systems*, Archive for Rational Mechanics and Analysis **58** (1975), no. 3, 181–205.

[Kee11] R. W Keener, *Theoretical statistics: Topics for a core course*, Springer, 2011.

[KHO97] A. Kohatsu-Higa and S. Ogawa, *Weak rate of convergence for an Euler scheme of nonlinear SDE's*, Monte Carlo Methods and Applications **3** (1997), 327–345.

[KLM08] S. J. Koopman, A. Lucas, and A. Monteiro, *The multi-state latent factor intensity model for credit rating transitions*, Journal of Econometrics **142** (2008), no. 1, 399–424.

[Kni00] K. Knight, *Mathematical statistics*, Texts in statistical science, Chapman & Hall/CRC, Boca Raton, Fla. ; London, 2000 (eng).

[Kor12] M. W. Korolkiewicz, *A dependent hidden Markov model of credit quality*, International Journal of Stochastic Analysis **2012** (2012).

[KP07] Y. M. Kaniovski and G. C. Pflug, *Risk assessment for credit portfolios: a coupled Markov chain model*, Journal of Banking & Finance **31** (2007), no. 8, 2303–2323.

[KP11] P. E. Kloeden and E. Platen, *Numerical Solution of Stochastic Differential Equations*, Stochastic Modelling and Applied Probability, Springer Berlin Heidelberg, 2011.

[KS01] A. Kreinin and M. Sidelnikova, *Regularization algorithms for transition matrices*, Algo Research Quarterly **4** (2001), no. 1/2, 23–40.

[KS12] I. Karatzas and S. Shreve, *Brownian motion and stochastic calculus*, Vol. 113, Springer, 2012.

[KS97] U. Küchler and M. Sorensen, *Exponential families of stochastic processes*, Vol. 3, Springer Science & Business Media, 1997.

[KŁSG02] M. Kostur, J. Łuczka, and L. Schimansky-Geier, *Nonequilibrium coupled Brownian phase oscillators*, Physical Review E **65** (2002), no. 5, 051115.

[KW13] A. Kremer and R. Weißbach, *Consistent estimation for discretely observed Markov jump processes with an absorbing state*, Statistical Papers **54** (2013), no. 4, 993–1007.

[KW14] A. Kremer and R. Weißbach, *Asymptotic normality for discretely observed Markov jump processes with an absorbing state*, Statistics & Probability Letters **90** (2014), 136–139.

[LC98] E. Lehmann and G Casella, *Theory of point estimation*, Springer-Verlag, 1998.

[LdRS15] A. Lionnet, G. dos Reis, and L. Szpruch, *Time discretization of FBSDE with polynomial growth drivers and reaction-diffusion PDEs*, Ann. Appl. Probab. **25** (2015), no. 5, 2563–2625. MR3375884

[Løf05] G. Løffler, *Avoiding the rating bounce: why rating agencies are slow to react to new information*, Journal of Economic Behavior & Organization **56** (2005), 365–381.

[Lin11] L. Lin, *Roots of stochastic matrices and fractional matrix powers*, Ph.D. Thesis, 2011.

[LR02] R. J. A. Little and D. B. Rubin, *Statistical analysis with missing data*, John Wiley & Sons, 2002.

[LS01] F. A Longstaff and E. S Schwartz, *Valuing American options by simulation: a simple least-squares approach*, The review of financial studies **14** (2001), no. 1, 113–147.

[LS02] D. Lando and T. M. Skodeberg, *Analyzing rating transitions and rating drift with continuous observations*, Journal of Banking and Finance **26** (2002), no. 2, 423–444.

[Lüt08] E. Lütkebohmert, *Concentration risk in credit portfolios*, Springer Science & Business Media, 2008.

[LV04] T. Lyons and N. Victoir, *Cubature on wiener space*, Proceedings of the royal society of london a: Mathematical, physical and engineering sciences, 2004, pp. 169–198.

[Mal03] F. Malrieu, *Convergence to equilibrium for granular media equations and their Euler schemes*, Ann. Appl. Probab. **13** (2003), no. 2, 540–560. MR1970276

[Mao08] X. Mao, *Stochastic differential equations and applications*, Horwood, 2008.

[MAT18] A. Matoussi, C. Alasseur, and I. B. Taher, *An extended mean field game for storage in smart grids* (2018).

[McK75] H. P. McKean, *Application of Brownian motion to the equation of Kolmogorov-Petrovskii-Piskunov*, Commun. Pure Appl. Math. **28** (1975), no. 3, 323–331.

[Mél96] S. Méléard, *Asymptotic behaviour of some interacting particle systems; Mckean-Vlasov and Boltzmann models*, Probabilistic models for nonlinear partial differential equations, 1996, pp. 42–95.

[MFE15] A. J McNeil, R. Frey, and P. Embrechts, *Quantitative risk management: Concepts, techniques and tools*, Princeton university press, 2015.

[MK07] G. McLachlan and T. Krishnan, *The EM algorithm and extensions*, Vol. 382, John Wiley & Sons, 2007.

[MPF12] D. S. Mitrinovic, J. Pecaric, and A. M. Fink, *Inequalities involving functions and their integrals and derivatives*, Vol. 53, Springer Science & Business Media, 2012.

[MS13] X. Mao and L. Szpruch, *Strong convergence and stability of implicit numerical methods for stochastic differential equations with non-globally lipschitz continuous coefficients*, Journal of Computational and Applied Mathematics **238** (2013), 14–28.

[MT05] G. Milstein and M. V Tretyakov, *Numerical integration of stochastic differential equations with nonglobally Lipschitz coefficients*, SIAM journal on numerical analysis **43** (2005), no. 3, 1139–1154.

[MW07] A. J. McNeil and J. P. Wendin, *Bayesian inference for generalized linear mixed models of portfolio credit risk*, Journal of Empirical Finance **14** (2007), no. 2, 131–149.

[Nea11] R. M Neal, *MCMC using Hamiltonian dynamics*, Handbook of Markov Chain Monte Carlo **2** (2011), no. 11, 2.

[Nor98] J. R. Norris, *Markov chains*, Cambridge university press, 1998.

[NP88] D. Nualart and É. Pardoux, *Stochastic calculus with anticipating integrands*, Probability Theory and Related Fields **78** (1988), no. 4, 535–581.

[NPV00] P. Nickell, W. Perraudin, and S. Varotto, *Stability of rating transitions*, Journal of Banking & Finance **24** (2000), no. 1, 203–227.

[Nua06] D. Nualart, *The malliavin calculus and related topics*, Vol. 1995, Springer, 2006.

[Oak99] D. Oakes, *Direct calculation of the information matrix via the EM*, Journal of the Royal Statistical Society: Series B (Statistical Methodology) **61** (1999), no. 2, 479–482.

[Pfe17] M. Pfeuffer, *ctmcd: An R Package for Estimating the Parameters of a Continuous-Time Markov Chain from Discrete-Time Data*, The R Journal (2017). To Appear.

[Pha09] H. Pham, *Continuous-time stochastic control and optimization with financial applications*, Vol. 61, Springer Science & Business Media, 2009.

[PP90] E. Pardoux and S. Peng, *Adapted solution of a backward stochastic differential equation*, Systems & Control Letters **14** (1990), no. 1, 55–61.

[PP92] É. Pardoux and S. Peng, *Backward stochastic differential equations and quasilinear parabolic partial differential equations*, Stochastic partial diff. equations and their applications (Charlotte, 1991), 1992, pp. 200–217.

[PR16] E. Pardoux and A. Râscanu, *Stochastic differential equations, Backward SDEs, Partial differential equations*, Springer, 2016.

[Pro05] P. E Protter, *Stochastic differential equations*, Springer, 2005.

[PRS18] M. Pfeuffer, G. d. Reis, and G. Smith, *Capturing Model Risk and Rating Momentum in the Estimation of Probabilities of Default and Credit Rating Migrations*, arXiv preprint arXiv:1809.09889 (2018).

[PT$^+$15] D. Possamaï, X. Tan, et al., *Weak approximation of second-order BSDEs*, The Annals of Applied Probability **25** (2015), no. 5, 2535–2562.

[RES18] G. d. Reis, S. Engelhardt, and G. Smith, *Simulation of McKean Vlasov SDEs with super linear growth*, arXiv preprint arXiv:1808.05530 (2018).

[RG15] C. H. Rhee and P. W. Glynn, *Unbiased estimation with square root convergence for SDE models*, Operations Research **63** (2015), no. 5, 1026–1043.

[Rob10] S. Robertson, *Sample path large deviations and optimal importance sampling for stochastic volatility models*, Stochastic Processes and their applications **120** (2010), no. 1, 66–83.

[RR98] S. T. Rachev and L. Rüschendorf, *Mass transportation problems. Vol. II*, Probability and its Applications (New York), Springer-Verlag, New York, 1998. Applications. MR1619171

[RRM10] A. Rasulov, G. Raimova, and M. Mascagni, *Monte Carlo solution of Cauchy problem for a nonlinear parabolic equation*, Mathematics and Computers in Simulation **80** (2010), no. 6, 1118–1123.

[RS18] G. d. Reis and G. Smith, *An unbiased Ito type stochastic representation for transport PDEs*, arXiv preprint arXiv:1804.03563 (2018).

[RST18] G. d. Reis, G. Smith, and P. Tankov, *Importance sampling for Mckean-Vlasov SDEs*, arXiv preprint arXiv:1803.09320 (2018).

[RT96] G. O Roberts and R. L Tweedie, *Exponential convergence of Langevin distributions and their discrete approximations*, Bernoulli **2** (1996), no. 4, 341–363.

[Sab13] S. Sabanis, *A note on tamed Euler approximations*, Electron. Commun. Probab. **18** (2013), no. 47, 10. MR3070913

[SBG04] B. Smith, P. Bjorstad, and W. Gropp, *Domain decomposition: parallel multilevel methods for elliptic partial differential equations*, Cambridge university press, 2004.

[SC11] J. Skoglund and W. Chen, *On the choice of liquidity horizon for incremental risk charges: are the incentives of banks and regulators aligned?*, The Journal of Risk Model Validation **5** (2011), no. 3, 37–57.

[SdR17] G. Smith and G. dos Reis, *Robust and consistent estimation of generators in credit risk*, 2017. To appear in Quantitative Finance, arXiv:1702.08867.

[Sko64] A. V. Skorokhod, *Branching diffusion processes*, Teor. Verojatnost. i Primenen. **9** (1964), 492–497.

[STT17] L. Szpruch, S. Tan, and A. Tse, *Iterative particle approximation for Mckean-Vlasov SDEs with application to multilevel Monte Carlo estimation*, 2017. ArXiv:1706.00907.

[Sup03] B. C. o. B. Supervision, *The new Basel capital accord* (2003).

[Sup09] B. C. o. B. Supervision, *Guidelines for computing incremental risk charge in the trading book* (2009).

[Sup13] B. C. o. B. Supervision, *Fundamental review of the trading book: A revised market risk framework* (2013).

[Szn91] A.-S. Sznitman, *Topics in propagation of chaos*, Ecole d'Eté de Probabilités de Saint-Flour XIX — 1989 (1991), 165–251.

[Szp10] L. Szpruch, *Numerical approximations of nonlinear stochastic systems*, Ph.D. Thesis, 2010.

[Tas04] D. Tasche, *The single risk factor approach to capital charges in case of correlated loss given default rates*, Available at SSRN 510982 (2004).

[TC03] H. Tsai and K. Chan, *A note on parameter differentiation of matrix exponentials, with applications to continuous-time modelling*, Bernoulli (2003), 895–919.

[Tes12] G. Teschl, *Ordinary differential equations and dynamical systems*, Vol. 140, American Mathematical Society Providence, RI, 2012.

[TFC16] H.-W. Teng, C.-D. Fuh, and C.-C. Chen, *On an automatic and optimal importance sampling approach with applications in finance*, Quantitative Finance **16** (2016), no. 8, 1259–1271.

[Tie13] D. N. Tien, *A stochastic Ginzburg-Landau equation with impulsive effects*, Physica A: Statistical Mechanics and its Applications **392** (2013), no. 9, 1962–1971.

[TÖ04] S. Trück and E. Özturkmen, *Estimation, adjustment and application of transition matrices in credit risk models*, Handbook of computational and numerical methods in finance, 2004, pp. 373–402.

[Tre00] L. N Trefethen, *Spectral methods in MATLAB*, Vol. 10, Siam, 2000.

[TW87] M. A. Tanner and W. H. Wong, *The calculation of posterior distributions by data augmentation*, Journal of the American statistical Association **82** (1987), no. 398, 528–540.

[Vil08] C. Villani, *Optimal transport: old and new*, Vol. 338, Springer Science & Business Media, 2008.

[VL78] C. Van Loan, *Computing integrals involving the matrix exponential*, Automatic Control, IEEE Transactions on **23** (1978), no. 3, 395–404.

[War17] X. Warin, *Variations on branching methods for nonlinear PDEs*, arXiv:1701.07660 (2017).

[War18] X. Warin, *Monte Carlo for high-dimensional degenerated Semi Linear and Full Non Linear PDEs*, arXiv preprint arXiv:1805.05078 (2018).

[Wat65] S. Watanabe, *On the branching process for Brownian particles with an absorbing boundary*, J. Math. Kyoto Univ. **4** (1965), 385–398.

[Wil67] R. Wilcox, *Exponential operators and parameter differentiation in quantum physics*, Journal of Mathematical Physics **8** (1967), no. 4, 962–982.

[Wil91] D. Williams, *Probability with martingales*, Cambridge university press, 1991.

[Wu83] C. F. J. Wu, *On the convergence properties of the EM algorithm*, The Annals of statistics (1983), 95–103.

[YM08] C. Yuan and X. Mao, *A note on the rate of convergence of the Euler-Maruyama method for stochastic differential equations*, Stochastic Analysis and Applications **26** (2008), no. 2, 325–333.

[YWZC14] T. Yavin, E. Wang, H. Zhang, and M. A. Clayton, *Transition probability matrix methodology for incremental risk charge*, Journal of Financial Engineering **1** (2014), no. 01.

[YZ99] J. Yong and X. Y. Zhou, *Stochastic controls: Hamiltonian systems and HJB equations*, Vol. 43, Springer Science & Business Media, 1999.

[Zei90] E. Zeidler, *Nonlinear functional analysis and its applications. II/B*, Springer-Verlag, New York, 1990. Nonlinear monotone operators, Translated from the German by the author and Leo F. Boron. MR1033498

[Zha04] J. Zhang, *A numerical scheme for BSDEs*, The Annals of Applied Probability **14** (2004), no. 1, 459–488.